Grundlagen der medizinischen

Biometrie

7. Auflage

mit Hyperlinks!

Hanns Ackermann

epsilon Verlag

Autor

Dr.rer.med. Dipl.-Math. Hanns Ackermann
Institut für Biostatistik & Math. Modellierung
Klinikum der Goethe-Universität Frankfurt
Theodor Stern - Kai 7
60590 Frankfurt am Main

Verlag

epsilon-Verlag GbR Darmstadt: 1. Aufl. 1995, 7. Aufl. 2014, eBook 2017, 2020

ISBN 3-9803214-6-0
ISBN 3-9806822-0-X

Modifizierte eBook-Versionen der 7. Auflage des gedruckten Buches 2017/2020:
Pdf-Datei mit Hyperlinks für Adobe Reader (1.5Mb, ab Pdf-Version 1.3 Acrobat 4.x)

Herstellung

Titeldesign: Schwarz auf Weiß, Darmstadt
Druck: Verlag Lindemann, Offenbach

Vorwort zur 7. Auflage

Das kann kein Zufall sein! Diese Feststellung trifft man fast täglich, aber: Ist es jetzt wirklich ein Zufall oder nicht? Eine Entscheidung fällt uns oft nicht leicht, und manchmal scheint es sogar unmöglich zu sein, im konkreten Fall zu einem richtigen Urteil zu kommen.

Statistische Methoden sollen uns bei solchen Entscheidungen helfen, nicht nur subjektiv, sondern vielmehr objektiv zu beurteilen. Eine erste Hilfe geben uns sogenannte Statistiken, die jedem bereits aus der Tagesschau geläufig sind, und so ist aber auch jedem geläufig, dass dabei ein breiter Interpretationsspielraum bleibt: Eine vielleicht gewünschte "Objektivität" bleibt dabei vielfach auf der Strecke, was speziell bei wissenschaftlichen Fragestellungen ganz sicher nicht der Fall sein sollte. Glücklicherweise geht aber nun die Mathematische Statistik weit über das hinaus, was volkstümlich so unter "Statistik" verstanden wird und schließt aus, dass subjektive Wünsche, subjektive Vorstellungen oder irgendwelche – gute oder auch weniger gute – Absichten in die Entscheidungsfindung eingehen. Die damit angesprochenen objektiven Methoden der wissenschaftlichen Statistik sind ausführlich Gegenstand dieses Skriptums.

Die Mathematische Statistik ging in ihrer Anwendung auf biologische und medizinische Fragestellungen eigene Wege, die durch die speziellen Probleme der Anwendungswissenschaften bedingt waren. Im Laufe der Zeit etablierte sich dadurch ein eigenes wissenschaftliches Fachgebiet, das heute im Allgemeinen als "Medizinische Biometrie" oder häufig auch als "Medizinische Statistik" bezeichnet und von eigenen Fachgesellschaften (zum Beispiel die Internationale Biometrische Gesellschaft, die Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie GMDS und andere) vertreten wird. In diesem Sinne sind die Inhalte dieses Skriptums als "Biometrische Methoden" zu verstehen, denn die Bezeichnung "Statistik" erinnert doch manchen allzu sehr an das, was eher subjektiven oder sogar manipulativen Charakter trägt.

Das vorliegende Skriptum entstand aus der Lehrtätigkeit des Autors am Universitätsklinikum Frankfurt, speziell aus der Vorlesung "Biomathematik für Mediziner", einer Biometrie-Vorlesung für Informatiker der TU Darmstadt mit Nebenfach Medizin und eines einführenden Biometrie-Kurses für Medizinisch-Technische Assistentinnen und Assistenten. Aus den ersten Arbeitsunterlagen zu diesen Veranstaltungen entwickelte sich im Laufe der Zeit die aktuelle Darstellung der *Grundlagen der Medizinischen Biometrie*.

Besonders für Ärztinnen und Ärzte, aber auch für Medizinstudierende mit kritischer Distanz zu MC-Fragen und Gegenstandskatalogen und mit eher lebenspraktischem Interesse an der Biometrie enthält der Text viele nützliche Hinweise und zahlreiche konkrete Beispiele zur Planung und Auswertung von eigenen klinischen Studien, stellt aber gleichzeitig auch eine ausreichende Grundlage für das biometrisch-statistische Verständnis bei der Lektüre der einschlägigen medizinischen Fachliteratur dar.

In diesem Skriptum steht neben den mathematischen und statistischen Grundlagen der Biometrie das inhaltliche Verständnis aller Methoden im Vordergrund, aus dem sich ohne jede "Rechnerei" das Verständnis der Ergebnisse einschlägiger PC-Programme ergibt: Alle Beispiele in diesem Text wurden mit dem ebenfalls im epsilon-Verlag erschienenen biometrischen Programmpaket *BiAS. für Windows* durchgerechnet und anhand der Programmausgaben interpretiert.

In Rücksichtnahme auf die sicher sehr unterschiedlichen mathematischen Vorbildungen und Leidenschaften potentieller Leserinnen und Leser wurden einige grundlegende mathematische Sachverhalte in einem Anhang zusammengestellt. Vielleicht dient dies den einen zur Auffrischung bereits erlernten Wissens, den anderen aber als erweiterbare Grundlage zum Verständnis der dargestellten Inhalte: Wer sich bereits vor dem Abitur von der Mathematik abgewendet hat, wird jedenfalls ermutigt, die Lektüre des Buches mit dem Lesen des Anhangs zu beginnen. Aber auch mathematisch eher abstinenten Leserinnen und Lesern ist ein Zugang zu der Materie möglich, da parallel zu allen vielleicht "problematischen" Passagen des Buches konsequent ein "formelfreier", mehr intuitiv gangbarer und von zahlreichen Beispielen und graphischen Illustrationen geleiteter Weg zum Verständnis des Stoffes verfolgt wird.

Ein Buch kann nicht ohne Hilfe anderer entstehen. Dr. Wolfgang Kirsten hat durch seine gründliche Lektüre und durch zahlreiche Anregungen zu diesem Skriptum beigetragen, gleichermaßen blieb die frühere Zeit der gemeinsamen Lehre mit Professor em. Klaus Abt nicht ohne Einfluss auf das Manuskript. Frau Professor Eva Herrmann und Professor em. André Kawerin danke ich sehr für ihr anhaltendes kritisches Interesse und für ihre immer hilfreichen Bemerkungen. Ganz herzlichen Dank auch an Frau Marion Kibbert-Ackermann und in memoriam an Frau Lydia Huck für ihre konstruktiven Beiträge! Und nicht zuletzt bin ich über die Jahre meinen Frankfurter Hörerinnen und Hörern zu großem Dank verpflichtet, denn an erster Stelle ist die erfolgreiche Wechselwirkung zwischen Lehrenden und Lernenden für die Gestaltung eines Lehrbuches entscheidend.

Die zweite bis siebte Auflage und die eBook-Version des Buches wurden fortlaufend überarbeitet, aktualisiert und, speziell zur sechsten Auflage, thematisch deutlich erweitert. Dies mit Dank an alle Leserinnen und Leser, die auf manche "Ecke und Kante" der laufenden Fassungen hingewiesen und durch zahlreiche Anregungen zur Verbesserung des Skriptums beigetragen haben!

Frankfurt am Main, im November 2014 und im Juli 2020 Hanns Ackermann

Inhalt

eBook: *Das Inhaltsverzeichnis enthält Hyperlinks! Ein Klick auf eine Überschrift führt zurück zum Inhaltsverzeichnis.*

Einleitung

Wozu benötigt man denn eigentlich in der Medizin statistische Methoden? Was haben denn Biologie und Medizin mit Mathematik und Statistik zu tun? Warum muss man sich als Ärztin oder Arzt, als MTA, in der Klinischen Chemie, in der Medizinischen Dokumentation oder auch in der Medizinischen Informatik mit Biometrie und Statistik beschäftigen?

Eine zufriedenstellende Antwort auf diese Fragen lässt sich sicher nicht ohne Weiteres in zwei Sätze fassen, wenn auch eine Motivation bereits in vielen alltäglichen Situationen erkennbar ist:

Nimmt man einen Gegenstand zur Hand und lässt ihn fallen, so fällt er bekanntlich immer zu Boden. Misst ein Arzt die Körperlänge eines erwachsenen Patienten, so wird er beim nächsten Besuch diese Messung kaum wiederholen, denn es würde sich wieder der gleiche Wert ergeben. Diese beiden nicht sehr spannenden Beispiele gewinnen aber im Vergleich mit einem anderen an Interesse: Beschäftigt man sich etwa mit dem Nüchternblutzuckerspiegel eines Patienten, so stellt man rasch fest, dass selbst wiederholte Messungen an der gleichen Person nicht zwangsläufig zu dem gleichen Resultat führen, denn im Gegensatz zu den ersten beiden Beispielen ist man bei dieser Untersuchung mit dem Phänomen der biologischen Variabilität konfrontiert:

Das Phänomen "Variabilität" beinhaltet, dass im individuellen Fall nicht immer das gleiche, identische Ergebnis zu beobachten ist, sondern dass man bei mitunter erheblichen Abweichungen nur "ungefähr" den gleichen Blutzuckerwert erhält. Biologische Variabilität findet man auch dann vor, wenn man nicht wiederholte Messungen an einer Person, sondern vielleicht jeweils eine Messung an vielen Personen vornimmt. Auch hier erhält man nicht immer identische Werte, sondern, bedingt durch die biologische Unterschiedlichkeit von Individuen, im Allgemeinen verschiedene Werte, die ebenfalls Ausdruck der biologischen Variabilität sind. "Im Mittel" (vorläufig mag man darunter den bekannten arithmetischen Mittelwert oder "Durchschnitt" verstehen) ist der Blutzuckerspiegel dagegen im Wesentlichen konstant: Eine Reproduzierbarkeit von Ergebnissen im Einzelfall darf man somit nicht erwarten, im Gegenteil gelingt eine Reproduzierung allenfalls *im Mittel* aller gleichartiger Messungen. Verfolgt man die Beispiele etwas weiter, so wird auch schnell klar, dass neben der Nicht-Konstanz eines Individuums auch mögliche Beobachtungs- bzw. Ablesefehler Ursache dafür sein können, dass man nicht immer identische Messwerte erhält. Bekanntlich spricht man dabei von einem "Messfehler", der seinerseits die Variabilität mitbedingen kann.

Hierzu lassen sich viele weitere, ebenso einfache Beispiele finden. Im medizinischen Bereich sind, von trivialen Ausnahmen abgesehen, kaum Messwerte zu finden, die nicht der biologischen Variabilität unterliegen, und so muss auch jeder Mediziner, jedes Labor und jede MTA mit dem

Phänomen "Biologische Variabilität" leben, und mehr noch, in allgemein akzeptierter Weise damit umgehen: Die Statistik gibt dazu Hilfestellung in Form einer Vielzahl von Methoden zur Beurteilung und zur Kontrolle der biologischen Variabilität, um damit eine nicht nur individuell-subjektive, sondern eine objektive Behandlung des beschriebenen Problems zu gestatten. Beispiele für solche Methoden finden sich reichlich in den nachfolgenden Kapiteln dieses Textes.

Neben den sogenannten *deskriptiven*, also nur beschreibenden statistischen Methoden existieren auch sogenannte *induktive* statistische Methoden, die vorliegende Daten nicht nur beschreiben, sondern darüber hinaus eine Grundlage dafür schaffen, weitergehende Schlüsse aus den Daten zu ziehen. Basis dieser Methoden ist die Überlegung, ob eine vorgefundene Situation noch mit dem Zufall erklärt werden kann oder ob sich – trotz Variabilität – in den Daten eine biologische oder medizinische Gesetzmäßigkeit erkennen lässt. Ist der Eindruck des englischen Arztes John Arbuthnot richtig, der 1710 feststellte, dass offenbar durchgängig pro Jahrgang mehr Jungen als Mädchen geboren werden? Muss man dieser Feststellung eine biologische Gesetzmäßigkeit unterstellen, oder kann die Beobachtung des Arztes, die sich ja nur auf einen gewissen Zeitraum bezieht, noch durch den Zufall (oder besser: durch die biologische Variabilität) erklärt werden? Sind – Stichwort Qualitätssicherung – möglicherweise vermutete Veränderungen der Resultate des Autoanalysers noch durch den Zufall im Sinne einer biologischen, technischen oder anderer Variabilität erklärbar? Ist das neue Immunsuppressivum tatsächlich erfolgreicher als der langjährige Standard, oder sind die gefundenen Unterschiede eher als unwesentlich zu beurteilen, denn irgendwelche, vielleicht "irrelevante", womöglich "zufällige" Unterschiede werden ja wohl letztlich in jeder Untersuchung zu Tage treten? Eine subjektive Beurteilung mag – hoffentlich – gut gemeint und vielleicht auch hilfreich sein, eine objektive, an den Gesetzmäßigkeiten des Zufalls orientierte, für jeden verbindliche Entscheidungsstrategie ist dabei sicher vorzuziehen.

Alles Beispiele für Mediziner, könnte man einwenden. Aber gerade deshalb auch für alle anderen Berufe im Umfeld der Medizin und des "Public-Health" relevant: Die berufliche Zusammenarbeit verlangt ein Verständnis für die grundlegenden Prinzipien der gängigen statistischen Methoden, die immer wieder in der medizinischen Forschung bei neuen therapeutischen und diagnostischen Fragestellungen, in der Laborroutine, bei der Qualitätssicherung, in der Medizinischen Dokumentation, beim Datenbankdesign und - nicht zuletzt! - bei der eigenen Lektüre von medizinischen Fachbüchern und -zeitschriften zum Tragen kommen. Dazu muss man kein Statistiker sein, sondern man sollte "nur" in der Lage sein, im aktiven wie passiven Umgang mit biometrisch-statistischen Methoden ein kritisches und kompetentes Verständnis zu entwickeln. Das vorliegende Skriptum entstand in der Hoffnung, seinen Leserinnen und Lesern dazu ausreichende Grundlagen zu vermitteln.

Kapitel 0: Wahrscheinlichkeitsrechnung

0.1 Wahrscheinlichkeiten

Die Wahrscheinlichkeitsrechnung trifft Aussagen über Ereignisse, die zufallsabhängig sind, also "nicht sicher" vorhersagbar sind - bekannte Beispiele sind das Würfeln und das Lotto-Spiel. In diesem Skript wird häufig von Wahrscheinlichkeiten gesprochen und es werden Begriffe wie "Irrtumswahrscheinlichkeit", "Wahrscheinlichkeitsverteilung" etc. verwendet, so dass eine Klärung des in der Statistik verwendeten Wahrscheinlichkeitsbegriffes erforderlich ist. Von einfachen Beispielen abgesehen wird im Rahmen dieses Kapitels noch kein besonderer Bezug zum biometrischen bzw. medizinischen Kontext hergestellt, so dass die notwendigen Begriffe zunächst auf eher abstrakter Ebene behandelt werden. Eine erste systematische Anwendung der Wahrscheinlichkeitsrechnung in der Medizin findet sich im 5. Abschnitt dieses Kapitels.

Um "Wahrscheinlichkeitsaussagen" treffen zu können, ordnet man jedem interessierenden Ereignis E eine Größe P(E) (i.e. eine *"Wahrscheinlichkeit"*, engl. *probability*) zu. Diese Zuordnung ist nicht einfach willkürlich vorzunehmen, sondern muss gewissen Bedingungen genügen, die 1933 von dem russischen Mathematiker Andrei Nikolajewitsch Kolmogoroff (1903-1987) axiomatisch formuliert wurden und in der heutigen Literatur als die drei *Kolmogoroffschen Axiome* bezeichnet werden:

I. Jedem Ereignis E ist eine Wahrscheinlichkeit P zwischen 0 und 1 zugeordnet: Es ist $0 \leq P(E) \leq 1$.

II. Das „sichere“ Ereignis S besitzt die Wahrscheinlichkeit 1: $P(S) = 1$.

III. Die Wahrscheinlichkeit, dass von zwei sich ausschließenden Ereignissen E_1 und E_2 das eine oder das andere eintritt, ist gleich der Summe der beiden Einzelwahrscheinlichkeiten: $E_1 \cap E_2 = \phi \Rightarrow P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

Das erste Axiom legt fest, dass Wahrscheinlichkeiten nicht negativ und nie größer als 1 sein können. Zum zweiten Axiom stelle man sich einen Topf mit Bohnen vor: Die Wahrscheinlichkeit, dass man beim Herausgreifen eines Gegenstandes eine Bohne zieht, ist gleich 1 (gefunden in einem Lehrbuch für Mathematische Statistik, Menges (1972)). Oder: Mit einem Würfel (jedenfalls mit einem Hexaeder!) wird man mit Sicherheit - mit Wahrscheinlichkeit 1 also - eine der Zahlen von 1 bis 6 würfeln, alles andere ist unmöglich. Aus dem dritten Axiom folgt, dass man bei Verletzung der Voraussetzung *"Disjunkte Ereignisse"* (Definition im Anhang!) die Wahrscheinlichkeiten nicht mehr ganz einfach addieren darf: Beispiele dazu finden sich weiter unten in diesem Kapitel.

Die drei Axiome definieren den Begriff "Wahrscheinlichkeit" indirekt, aber sehr allgemein durch Angabe wünschenswerter Eigenschaften, während ein vielleicht eher geläufiger, direkter Zugang bereits im 17. Jahrhundert durch Jakob Bernoulli (1655-1705) formuliert wurde ("ars conjectandi", 1713 posthum publiziert von seinem Enkel Nikolaus Bernoulli). Prinzipiell unterscheidet Bernoulli sogenannte *a-priori*- und *a-posteriori*-Definitionen:

Der *klassische* oder auch *logische Wahrscheinlichkeitsbegriff*, vertreten z.B. von Pierre-Simon de Laplace (1749-1827), definiert - ohne empirisches Wissen - die a-priori-Wahrscheinlichkeit für das Ereignis E deduktiv durch

$$P(E) = \frac{\text{Anzahl günstiger Ereignisse}}{\text{Anzahl möglicher Ereignisse}}$$

Wirft man eine Münze in die Luft, so gibt es - spitzfindige Einwände ausgeschlossen – zwei mögliche Ausgänge: Kopf oder Zahl. In der Definition von P(E) gibt der Zähler die Anzahl Möglichkeiten für zum Beispiel "Kopf" an (hier:1), der Nenner entspricht der Anzahl überhaupt möglicher Ereignisse, im Beispiel also 2. Die Wahrscheinlichkeit P(E) ist damit gleich 1/2.

Beim Würfeln mit *einem* Würfel erhält man vermöge nur genau einer von 6 Möglichkeiten eine "3": Die Wahrscheinlichkeit, eine "3" zu würfeln, beträgt damit 1/6. Würfelt man mit *zwei* Würfeln, bildet die Augen*summe* und fragt sich nach der Wahrscheinlichkeit, die Augen*summe* "3" zu würfeln, so wird es etwas komplizierter: Jeder der beiden Würfel kann 6 Augenzahlen zeigen, somit gibt es $6 \cdot 6 = 36$ *Elementarereignisse* (Nenner!). Zwei davon führen zur Augensumme "3", nämlich 1+2=3 und 2+1=3. Damit ist die gesuchte Wahrscheinlichkeit gerade 2/36. Mit etwas Geduld kann man dieses Ergebnis ohne Weiteres mit Hilfe der empirischen Definition überprüfen:

Im Beispiel des Münzwurfs stellt man nach wiederholten Versuchen im Laufe der Zeit fest, dass die beiden *relativen Anteile* von "Kopf" und "Zahl" immer stabiler werden, sich immer mehr ausgleichen und, falls man bis ans Ende aller Tage weiterwirft, vermutlich irgendwann gleich sind: Der Anteil des Ereignisses "E=Kopf" (oder auch "E=Zahl") stimmt mit der größer werdenden Anzahl n der Versuche immer besser mit einem Grenzwert überein, den man als *empirische* oder auch als *frequentistische Wahrscheinlichkeit* bezeichnet:

$$P(E) = \lim_{n \to \infty} \frac{\text{Anzahl zu „E“ führender Versuche}}{\text{Gesamtanzahl n der Versuche}}$$

Ein bekannter Verfechter dieser a-posteriori-Auffassung ist der Mathematiker Richard von Mises (1883-1953). Den Quotienten rechts vom limes-Zeichen (lim=limes=Grenzwert) bezeichnet man auch als *relative Häufigkeit*, die - wie auch die Wahrscheinlichkeit - offensichtlich nur Werte zwischen 0 und 1 annehmen kann.

Bitte beachten Sie, dass die beiden Definitionen nicht zwangsläufig zum gleichen numerischen Ergebnis führen, wie das Beispiel von Mädchen- und Knabengeburten zeigt: *Logisch* bzw. *a-priori* ist P(Mädchen)=P(Junge), *frequentistisch* dagegen ist P(Mädchen)<P(Junge)!

0.2 Rechenregeln für Wahrscheinlichkeiten

Die bekanntesten Regeln für das Rechnen mit Wahrscheinlichkeiten sind der *Additionssatz* und der *Multiplikationssatz*. Man unterscheidet dabei zwischen sogenannten *vereinbaren* und *unvereinbaren* Ereignissen: Unter *vereinbaren Ereignissen* versteht man Ereignisse, die auch gleichzeitig bzw. gemeinsam auftreten können (z.B. *unabhängige Ereignisse*), unter *unvereinbaren Ereignissen* solche, die sich gegenseitig ausschließen.

Beim Würfeln schließen sich die beiden Ereignisse A="Augenzahl=1" und B="Augenzahl≥5" gegenseitig aus und sind deshalb *un*vereinbar. Mit dem Additionssatz für unvereinbare Ereignisse kann die Wahrscheinlichkeit für das Ereignis E, eine "1" (Ereignis A) oder andererseits "5 oder 6" (Ereignis B) zu würfeln errechnet werden (3. Kolmogoroffsches Axiom, das aus der Mengenlehre bekannte Vereinigungszeichen "$\cup$" weist auf die mengentheoretische Interpretation hin, vgl. Anhang A.2); im Beispiel erhält man das Ergebnis 1/6+2/6=1/2.

$P(E) = P(A \text{ oder } B) = P(A \cup B) = P(A) + P(B)$ (A und B unvereinbar)

Die möglichen Ergebnisse beim Würfeln mit *zwei* Würfeln sind in Abbildung 1 dargestellt, die hellen Felder bedeuten die Augensumme:

Σ	**1**	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Abbildung 1: Würfelschema für Augensummen

Wie groß ist nun die Wahrscheinlichkeit, beim Würfeln mit *zwei* Würfeln *zwei* Sechsen zu erhalten? Abbildung 1 unterscheidet dazu 6·6=36 Möglichkeiten (*Elementarereignisse*), eine davon führt zur Augensumme 12: P(Σ=12)=1/36. Oder: Die Chance, mit dem ersten Würfel eine "6" zu würfeln ist 1/6=6/36 (Ereignis A, letzte Spalte), in 1/6 dieser Fälle wird auch mit dem zweiten Würfel eine "6" gewürfelt (Ereignis B, letzte Zeile), also wird in 1/6 von 1/6 aller Fälle mit beiden Würfeln eine "6" gewürfelt. Allgemeiner bedeutet dies die Wahrscheinlichkeit für das gemeinsame Auftreten von unabhängigen Ereignissen:

$P(E) = P(A \text{ und } B) = P(A \cap B) = P(A) \cdot P(B)$ (A und B unabhängig)

Mit Hilfe dieser Definition des *Multiplikationssatzes für unabhängige Ereignisse* kann man nun auch den *Additionssatz für vereinbare, auch unabhängige Ereignisse* formulieren. Wie groß ist die Wahrscheinlichkeit, mit dem ersten *oder* mit dem zweiten *oder* mit beiden Würfeln eine Sechs zu würfeln (*"einschließendes Oder"*)? Nach dem Additionssatz für unvereinbare Ereignisse könnte man zunächst vermuten, dass diese Wahrscheinlichkeit 1/6+1/6 beträgt. Nun beinhalten aber *beide* (vereinbaren!) Ereignisse das Ergebnis "beide Würfel 6" (*beide* Würfel können zu einer "6" führen!): In der einfachen Summe der beiden Einzelwahrscheinlichkeiten wird die Wahrscheinlichkeit $P(A \cap B)$ für das Ereignis "beide Würfel 6" (also $P(A \cap B) = P(1.\text{Würfel}=6) \cdot P(2.\text{Würfel}=6)$) demnach doppelt berücksichtigt und muss somit wieder einmal abgezogen werden. Der *Additionssatz* lautet also allgemein für vereinbare, speziell auch für unabhängige Ereignisse

$P(E) = P(A \text{ oder } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (A und B vereinbar)

In einem medizinischen Beispiel: Die Wahrscheinlichkeit für Erkrankung A sei P(A)=0.10, die für eine von A unabhängige Erkrankung B sei P(B)=0.02. Die Wahrscheinlichkeit, im Laufe der Zeit an A oder B zu erkranken, ist zunächst "ungefähr" P(A)+P(B)=0.10+0.02. Da man dabei aber alle Personen, die sowohl an A als auch an B erkranken, in Bezug auf A *und* B berücksichtigt, muss man im Gesamtresultat die Wahrscheinlichkeit für "sowohl A als auch B" (vgl. Sie dazu bitte den Multiplikationssatz!) subtrahieren; aus der letzten Formel ergibt sich $0.118 = 0.10 + 0.02 - 0.10 \cdot 0.02$. Die mengentheoretische Schreibweise erleichtert das Verständnis: Stellt man alle Personen mit A in einer Menge/Gruppe zusammen und macht das gleiche mit allen Personen mit Erkrankung B, so gibt es zwangsläufig eine Schnittmenge derjenigen, die sowohl zur Menge/Gruppe A als auch zu B gehören: Die einfache Summe der Anteile ist somit zu groß, denn man hat die an A Erkrankten, die auch an B leiden, in *beiden* Gruppen mitgezählt. Da A und B unabhängig sind, erkranken offenbar 2% aller an A Erkrankten auch an B, das sind, auf die Gesamtbevölkerung bezogen, 2% von 10% aller Menschen, also das Produkt $2\% \cdot 10\% = 0.02 \cdot 0.10 = 0.002 = 0.2\%$. Diese Zahl wiederum von 0.120 subtrahiert, ergibt die bereits oben errechnete Wahrscheinlichkeit $P(A \cup B) = 0.118$.

0.3 Bedingte Wahrscheinlichkeiten

Unter der bedingten Wahrscheinlichkeit eines Ereignisses A versteht man die Wahrscheinlichkeit von A unter der Bedingung, dass bereits ein weiteres Ereignis B eingetreten ist, formal:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Mit dieser Beziehung kann man den oben formulierten Multiplikationssatz für unabhängige Ereignisse - für diese ist offenbar $P(A|B)=P(A)$ – etwas allgemeiner fassen: Die Wahrscheinlichkeit für das gleichzeitige Auftreten

zweier unabhängiger oder abhängiger, vereinbarer oder unvereinbarer Ereignisse A und B beträgt, wegen $(A \cap B)=(B \cap A)$ in symmetrischer Definition,

$$P(A \cap B) = P(B) \cdot P(A \mid B) = P(A) \cdot P(B \mid A) = P(B \cap A)$$

Bedingte Wahrscheinlichkeiten treten auf, wenn man von der Wahrscheinlichkeit eines Ereignisses A nur unter bestimmten Umständen sprechen möchte. Die Schreibweise P(A|B) bedeutet die Wahrscheinlichkeit für A, vorausgesetzt, B ist bereits eingetreten.

Es leiden mehr Männer als Frauen an Gicht, so dass man eine Beziehung zwischen den beiden zugehörigen bedingten Wahrscheinlichkeiten formulieren kann: Es ist P(Gicht|Mann)> P(Gicht|Frau). Oder: Aus genetischen Gründen unterscheiden sich die Anteile der Blutgruppen A, B, AB und 0 in verschiedenen Ländern, z.B. ist P(0|Franzose) ≠ P(0|Schweizer).

Aus dem Begriff der bedingten Wahrscheinlichkeit kann man das *Relative Risiko* ableiten. Menschen, die einem bestimmten Risikofaktor R ausgesetzt sind, besitzen eine andere Wahrscheinlichkeit, an einer bestimmten Krankheit K zu erkranken als Nicht-Exponierte ($\overline{R}$, der Querstrich bedeutet stets das Gegenteil bzw. Komplement eines Ereignisses oder, wie hier, einer Bedingung R). Die erste dieser beiden Wahrscheinlichkeiten sei $P(K|R)$, die zweite $P(K|\overline{R})$. Das Relative Risiko RR ist nun definiert durch

$$RR = \frac{P(K|R)}{P(K|\overline{R})}$$

Das dem Risikofaktor R *zuschreibbare Risiko* δ ist definiert durch die Differenz der beiden Wahrscheinlichkeiten $P(K|R)$ und $P(K|\overline{R})$:

$$\delta = P(K|R) - P(K|\overline{R})$$

0.4 Das Bayessche Theorem

Nach der Definition der bedingten Wahrscheinlichkeit erhält man aus dem Multiplikationssatz (mittlere Terme!)

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

Zerlegt man nun das Ereignis B in zwei *disjunkte* Teilmengen gemäß $B=(A \cap B) \cup (\overline{A} \cap B)$ und wendet die allgemeine Form des Multiplikationssatzes (1. und 3. Term!) auf die beiden Wahrscheinlichkeiten $P(A \cap B)$ und $P(\overline{A} \cap B)$ an, so erhält man $P(B)=P(A) \cdot P(B|A)+P(\overline{A}) \cdot P(B|\overline{A})$ als Summe der Wahrscheinlichkeiten zweier disjunkter Ereignisse. Damit ergibt sich aus der letzten Beziehung die bekannte *Bayessche Formel*, die, publiziert posthum 1763, auf Reverend Thomas Bayes (1702-1761) zurückgeht:

$$P(A|B) = \frac{P(A)\cdot P(B|A)}{P(A)\cdot P(B|A) + P(\overline{A})\cdot P(B|\overline{A})}$$

Es sei zum Beispiel A eine bestimmte Erkrankung und B ein Symptom. Ist die a-priori-Wahrscheinlichkeit P(A) bekannt und weiß man, mit welchen Wahrscheinlichkeiten das Symptom bei Krankheit bzw. bei Nicht-Erkrankung beobachtet wird (das sind die beiden bedingten Wahrscheinlichkeiten), so kann man vermöge der Bayesschen Formel die a-posteriori-Wahrscheinlichkeit P(A|B) ausrechnen: Dies ist die Wahrscheinlichkeit, dass bei festgestelltem Symptom B die Erkrankung A vorliegt. (Cave: Die beiden Begriffe "a-priori" und "a-posteriori" werden hier *nicht* im Sinne von Abschnitt 0.1 verwendet!)

Ein diagnostischer Test T soll zum Nachweis einer Krankheit K verwendet werden. Setzt man in der Bayesschen Formel A=K und B=T^+, so bedeutet P(K|T^+) die Wahrscheinlichkeit, dass bei einem positiven Testergebnis tatsächlich die Erkrankung K vorliegt. Die Wahrscheinlichkeit P(T^+|K) erhält man, indem man den Test bei Patienten anwendet, bei denen die fragliche Erkrankung definitiv feststeht und prüft, ob der Test positiv ausfällt. Testet man Personen, die nicht an K leiden, so erhält man auf dem gleichen Weg die Wahrscheinlichkeit P(T^+|$\overline{K}$). Bei bekannter a-priori-Wahrscheinlichkeit P(K) errechnet man mit der Bayesschen Formel die gewünschte Wahrscheinlichkeit P(K|T^+). Mehr dazu in Abschnitt 6.4.

Die Bayessche Formel kann verallgemeinert werden, wenn nicht nur zwei Ereignisse A und $\overline{A}$, sondern n sich ausschließende Ereignisse A_i (zum Beispiel n Diagnosen) bei gegebener Bedingung (Symptomkomplex) B untersucht werden sollen; in der Bayesschen Formel ersetzt man A durch A_k ($1 \le k \le n$) und den Nenner durch $\Sigma_i(P(A_i)\cdot P(B|A_i))$, um die bedingte Wahrscheinlichkeit $P(A_k|B)$ des Ereignisses A_k (der Diagnose A_k) zu erhalten:

$$P(A_k|B) = \frac{P(A_k)\cdot P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)\cdot P(B \mid A_i)}$$

0.5 Die Binomialverteilung

Ohne großen Aufwand ergibt sich aus dem Stand der Definitionen aus Abschnitt 0.2 eine erste Wahrscheinlichkeitsverteilung, die sogenannte *Binomialverteilung*. Die Binomialverteilung, gelegentlich auch nicht ganz korrekt als *Bernoulli-Verteilung* bezeichnet, befasst sich mit zwei alternativen Ereignissen E ("Erfolg", z.B. "Kopf") und $\overline{E}$ ("Misserfolg", im Beispiel entsprechend "Zahl"); die Wahrscheinlichkeit für einen "Erfolg" sei $P(E)=\theta$ und die komplementäre Wahrscheinlichkeit für einen "Misserfolg" sei $P(\overline{E})=1-\theta$. In Frage steht die Wahrscheinlichkeit, in einem Experiment bei n Versuchen k Erfolge zu haben:

Wie groß ist die Wahrscheinlichkeit, beim Münzwurf k=5 mal "Kopf" zu erhalten, wenn die Münze n=10 mal geworfen wird? Wie groß ist die Wahrscheinlichkeit, dass unter den n=4 Kindern einer Familie genau ein Mädchen (k=1) oder mindestens ein Mädchen (k≥1) ist? Die gesuchte Wahrscheinlichkeit, bei n Versuchen genau k Erfolge und n-k Misserfolge zu haben, ergibt sich mit Hilfe des Multiplikations- und des Additionssatzes für Wahrscheinlichkeiten:

Im Beispiel der n=4 Kinder - darunter k=1 Mädchen (E) und n-k=3 Jungs ($\bar{E}$) - ist die Wahrscheinlichkeit für irgendeine bestimmte Reihenfolge der Ergebnisse "E $\bar{E}$ $\bar{E}$ $\bar{E}$" gegeben durch das Produkt $\theta\cdot(1-\theta)\cdot(1-\theta)\cdot(1-\theta)$, oder, äquivalent, $\theta^1\cdot(1-\theta)^3=\theta^1\cdot(1-\theta)^{4-1}=\theta^k\cdot(1-\theta)^{n-k}$ (Multiplikationssatz). Das Gleiche gilt für jede andere Reihenfolge der E und $\bar{E}$, Analoges auch für beliebige n und k.

Im Beispiel von n=4 und k=1 gibt es vier unterscheidbare Elementarereignisse, dass nämlich das Mädchen das erste, das zweite, etc. Kind ist, wobei *jede* dieser vier Varianten die oben definierte Wahrscheinlichkeit $\theta^1\cdot(1-\theta)^3$ besitzt. Im allgemeinen Fall von k>1 errechnet man mit Hilfe des Binomialkoeffizienten $\binom{n}{k}$ (Anhang A.3!) die Anzahl der unterschiedlichen Anordnungen der beiden möglichen Ereignisse "Junge" und Mädchen". Vermöge des Additionssatzes für disjunkte Ereignisse definiert sich damit summarisch die *Binomialverteilung*:

$$P_{\theta,n,k} = \binom{n}{k} \cdot \theta^k \cdot (1-\theta)^{n-k}$$

Die Binomialverteilung kann graphisch in Form eines diskret skalierten Diagramms dargestellt werden, womit auch im Wortsinn der Begriff *Wahrscheinlichkeitsverteilung* anschaulich wird:

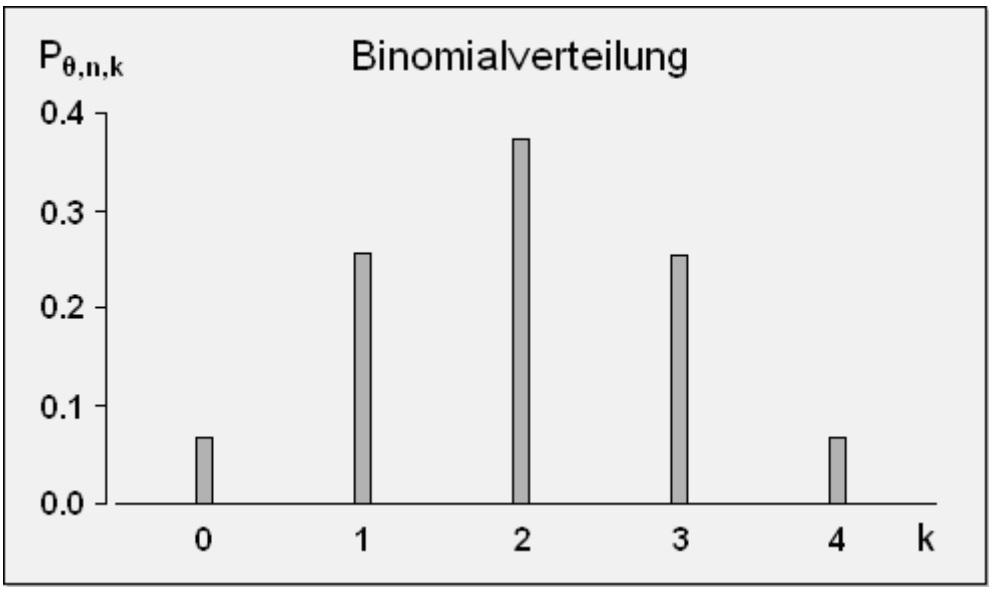


Abbildung 2: Binomialverteilung für θ=0.5 und n=4

Die Wahrscheinlichkeit, dass man bei n Versuchen nicht genau, sondern *mindestens* k Erfolge erzielt, kann somit durch die *Summen-* oder *Verteilungsfunktion* der Binomialverteilung berechnet werden:

$$B_{\theta,n,k} = \sum_{i=k}^{n} \binom{n}{i} \cdot \theta^i \cdot (1-\theta)^{n-i}$$

Die Wahrscheinlichkeit für mindestens ein Mädchen unter 4 Kindern ergibt sich daraus mit $B_{0.5,4,1}=0.9375$. Die Wahrscheinlichkeit, nur Jungs zu haben ist demnach $1-B_{0.5,4,1}=1-0.9375=0.0625$. "Ziemlich unwahrscheinlich, kommt aber mal vor" könnte man die Sache abtun. Bei 5 Kindern ist man etwas unsicherer, denn da ist $1-B_{0.5,5,1}=0.03125$; bei 5 Kindern ist die Wahrscheinlichkeit, nur Jungen zu haben, deutlich kleiner als bei vier Kindern. Wenn man aber nun behauptet: "Das geht nicht mit rechten Dingen zu!" begeht man möglicherweise einen Fehler, denn eine Familie mit dieser Konstellation ist ja durchaus denkbar, mit einer Wahrscheinlichkeit von eben p=0.03125 (rund 31 von 1000 Familien mit 5 Kindern haben 5 Jungs, aber keine Mädchen): Die damit angesprochene statistische Prüfung von Hypothesen wird ausführlich Thema des 5. Kapitels sein.

Die Inzidenzrate akuter Leukämien bei ≤4-jährigen wird in der BRD mit $\theta=0.000104$ angegeben (Quelle: Keller, Haaf, Kaatsch, Michaelis (1990), Untersuchung von Krebserkrankungen im Kindesalter in der Umgebung westdeutscher kerntechnischer Anlagen, *Bundesministerium UNR*, vgl. dazu auch Kaatsch et al. (2008), *Deutsches Ärzteblatt International* 105(42): 725–32). Im Umkreis von 30 km um ein Atomkraftwerk leben n=4000 Kinder im Alter von bis zu 4 Jahren. Innerhalb eines Jahres werden k=3 neue Fälle von akuter Leukämie erfasst (außer θ fiktive Daten). Mit Hilfe der Binomialverteilung errechnet man dazu die Wahrscheinlichkeiten $P_{\theta,4000,3}=0.0079$ (genau drei Fälle) und $B_{\theta,4000,3}=0.0088$ (drei oder mehr Fälle): Die Wahrscheinlichkeit, dass bei unterstellter identischer Inzidenzrate drei oder mehr Kinder innerhalb eines Jahres im fraglichen Gebiet erkranken, beträgt 0.0088. Erwartet hätte man $\theta \cdot n = 0.000104 \cdot 4000 = 0.416000$ Erkrankungsfälle pro Jahr (bzw. etwa einen Erkrankungsfall in 2.5 Jahren). Auf eine formale Bewertung dieser Ergebnisse wird in Abschnitt 5.1 im Rahmen der statistischen Nullhypothesenprüfung näher eingegangen.

Die bei gegebener Wahrscheinlichkeit θ und gegebener Anzahl Versuche n *erwartete Anzahl* der "Erfolge" wird als *Erwartungswert* μ bezeichnet; im Falle der Binomialverteilung ist $\mu=\theta \cdot n$. Die *Varianz* der Binomialverteilung beträgt $\sigma^2=\theta \cdot (1-\theta) \cdot n$, die *Standardabweichung* σ als Wurzel aus σ^2 ist die "mittlere" bzw. "erwartete" Abweichung vom Erwartungswert μ. Die genannten Begriffe werden im nächsten Absatz anhand eines numerischen Beispiels und ausführlicher in Abschnitt 3.2 dargestellt.

Im Beispiel der Leukämieinzidenzen beträgt der Erwartungswert μ der Wahrscheinlichkeitsverteilung der Leukämie-Neuerkrankungen $\mu=\theta \cdot n=0.000104 \cdot 4000=0.416000$, die Varianz σ^2 ist $\theta \cdot (1-\theta) \cdot n=0.000104 \cdot 0.999896 \cdot 4000=0.415957$ und die Standardabweichung $\sigma=\sqrt{\sigma^2}$ beträgt damit $\sigma=0.644947$. Im Mittel aller *gleichartigen* Untersuchungen *erwartet* man bei der angenommenen Inzidenzrate von $\theta=0.000104$ demnach die Anzahl $\mu=0.416000$ von Neuerkrankungen pro Jahr, die erwartete ("mittlere" oder "durchschnittliche") Abweichung von dem an sich erwarteten Wert μ beträgt "von Studie zu Studie" gerade $\sigma=0.644947$.

Kapitel 1:
Skalen, Daten und Datengewinnung

In der Planungsphase einer wissenschaftlichen Untersuchung befasst man sich als Erstes mit einer präzisen Charakterisierung aller relevanten Daten, wie im ersten Abschnitt dieses Kapitels ausführlich dargestellt wird. Der zweite Abschnitt befasst sich mit den Unterschieden zwischen Datenerhebungen und Experimenten bzw. Versuchen, denn auch davon hängt ganz entscheidend ab, welche statistischen Methoden zur Auswertung heranzuziehen sind. Der dritte Abschnitt gibt einen Einblick in die klassische Methodik der statistischen Studienplanung und -auswertung.

1.1 Statistische Skalen und Daten

Die Statistik unterscheidet zwischen *quantitativen* und *qualitativen* Skalen, die wiederum *stetiger* oder *diskreter* Natur sein können:

Alle Daten, die man messen oder zählen kann, werden als *quantitative Daten* bezeichnet, und entsprechend ist die Skala, bezüglich derer die Daten bestimmt werden, eine *quantitative Skala;* Beispiele dazu sind das Schlagvolumen des Herzens, Strecken im EKG, Knochendichte etc.. Quantitative Daten lassen sich in der Regel durch reelle oder natürliche Zahlen ausdrücken. Gelegentlich unterscheidet man auch Skalen mit und ohne absoluten Nullpunkt, dies sind *Absolutskalen* und *Intervallskalen*.

Nicht-quantitative Daten werden als *qualitativ* bezeichnet, die entsprechenden Skalen heißen ebenfalls *qualitativ*. Qualitative Daten besitzen keine zahlenmäßigen Wertbeimessung und sind nicht quantifizierbar, insbesondere darf man mit solchen Daten keine Rechenoperationen wie Addition und Multiplikation durchführen. Einfache Beispiele - Blutgruppe, Rhesusfaktor, Hernienlokalisation oder Schmerzgrad - machen dies ohne Weiteres verständlich. Kann man, wie beim Schmerzgrad, die Daten in eine *Rangfolge* bringen, so spricht man auch von *ordinalen Daten* (gelegentlich auch von *Rangdaten*), ist dies wie in den anderen Beispielen nicht möglich, so spricht man von *Nominaldaten* oder von *kategorialen* Daten. Die zugehörigen Skalen bezeichnet man als *Ordinalskalen* oder *Rangskalen* bzw. als *Nominalskalen*. Eine häufig vorgenommene Nummerierung der Stufen einer Rangskala (Beispiel Schmerz: 0,1,2,3 an Stelle von kein, leichter etc. Schmerz) bedeutet *keine* "Quantifizierung", sondern lediglich eine Umbenennung! Man spricht bei solchen Rangnummern auch von *Ordinalzahlen*, mit denen man aber bekanntlich nicht rechnen kann.

Eine Skala kann *stetig* oder *diskret* sein. Als Beispiel für diskrete Daten bzw. Zahlen kann man die natürlichen Zahlen nennen, während die reellen

Zahlen von ihrer Natur her stetig sind (Anhang A.1). "Diskret" bedeutet also, dass man nur "gewisse" Werte erhalten kann (Anzahl Kinder pro Familie!), während eine stetige Skala jeden beliebigen Wert zulässt. Stetige Skalen findet man etwa bei Körpergewicht, Herzfrequenz und Blutdruck vor, diskrete Skalen bei Kinderzahl, Geschlecht, Diagnose u.a.. Als Synonym für *stetig* wird oft auch die Bezeichnung *kontinuierlich* verwendet.

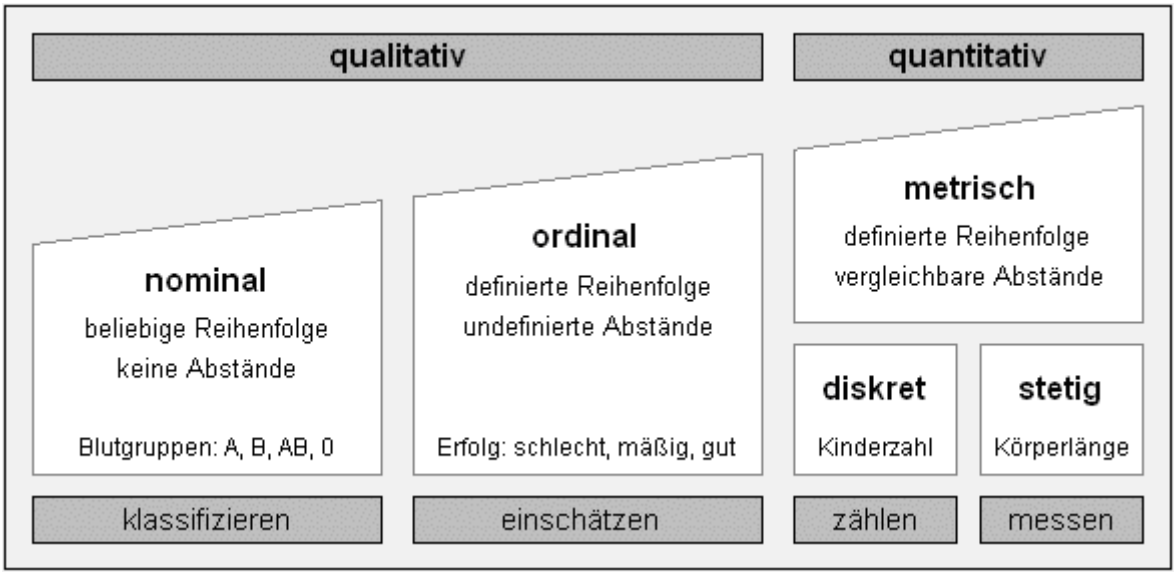


Abbildung 3: Skalen

Die Unterscheidung von stetigen und diskreten Skalen ist wider Erwarten sogar bei qualitativen Skalen von Bedeutung: Im Beispiel des Schmerzgrades verwendet man gelegentlich ein "Dolometer", das realiter nichts anderes ist als eine Strecke bzw. ein reelles Intervall von 0 bis 1 bzw. von 0% bis 100%. Bittet man einen Patienten oder Probanden, auf irgendeine Stelle auf dieser Skala zu zeigen, die seinem subjektiven Schmerzempfinden entspricht, so erhält man - trotz qualitativer Skala - als Resultat ein stetiges Datum. An Stelle von "Dolometer" oder Ähnlichem wird oft auch die Bezeichnung *Visuelle Analogskala* (VAS) verwendet.

Pragmatisch betrachtet, kann in der Praxis nie von stetigen Skalen gesprochen werden, denn bei allen Laborwerten, physikalischen Messungen etc. kann man – insbesondere bedingt durch die Messgenauigkeit - nur eine "gewisse", endliche Anzahl Stellen einer Zahl zu Papier bringen: Damit hat man eine an sich stetige Skala *künstlich diskretisiert*. Wie sich später herausstellt, können statistische Methoden, die ursprünglich für stetige Skalen konzipiert wurden, trotzdem in aller Regel auch auf "künstlich diskretisierte Daten" angewendet werden.

In den oben verwendeten Beispielen der Variablen Körpergröße, Haarfarbe etc. spricht man auch von sogenannten *Merkmalen* (engl.: *endpoint*) und bezeichnet damit diejenigen Größen, über die man eine Aussage machen möchte. Merkmale werden an *Merkmalsträgern* beobachtet oder gemessen; Merkmalsträger (synonym: *Beobachtungseinheiten*) sind Patienten, Probanden, Versuchstiere, Petrischalen oder anderes. Die konkrete Messung bzw. Beobachtung wird auch als *Merkmalsausprägung* bezeichnet.

1.2 Datenerhebungen, Experimente, Studien

Die Bezeichnung "Studie" wird üblicherweise sowohl für Datenerhebungen als auch für medizinische Experimente bzw. Versuche verwendet. Die klassische Datenerhebung ist sehr der früher so genannten "*Medizinalstatistik*" verwandt, während die statistische Methodik *epidemiologischer und kontrollierter klinischer Studien* (*Controlled* oder *Randomized Clinical Trials*, CCT/RCT) eindeutig der modernen Biometrie zuzurechnen ist:

Datenerhebungen untersuchen vorhandene Zustände. Man unterscheidet dabei zwischen *retrospektiven* und *prospektiven* Erhebungen: Bei retrospektiven Studien verwendet man bereits vorhandene Daten (z.B. Archivdaten der letzten zehn Jahre o.ä.) und versucht, damit seine Fragen zu beantworten. Retrospektive Studien bringen in aller Regel Probleme mit sich, denn oft sind die fraglichen Daten unvollständig, vielleicht ungenau und/oder unsystematisch geführt, oder, sicher das oft entscheidendste Problem, möglicherweise nach einem unbekannten Mechanismus selektiert. Eine *Selektion* (z.B. eine Altersselektion von Patienten) kann zu einer Verfälschung aller Aussagen führen und kann damit eine Studie weitgehend wertlos machen. Einen Schutz gegen solche - in der Regel unbekannten - Selektionsmechanismen gibt es nicht.

Eine *prospektive Studie* beginnt zu einem definierten Zeitpunkt und erfasst von da an alle interessierenden Größen nach einem festen Plan, der auch im Laufe der Erhebung nicht geändert wird. Zwar ist eine prospektive Studie im Allgemeinen mit größerem Aufwand, vielleicht auch mit höheren Kosten verbunden, trotzdem sollte diese Form der retrospektiven Variante vorgezogen werden, zumal die Vorteile offensichtlich sind.

Gelegentlich ist es von Interesse, nicht nur zu einem bestimmten Zeitpunkt eine "Bestandsaufnahme" zu machen, sondern über einen gewissen Zeitraum ein *Kollektiv* (eine "*Kohorte*") von Patienten oder Probanden zu beobachten. Im ersten Fall spricht man von einer *Querschnittstudie*, im zweiten Fall von einer *Längsschnittstudie*. Eine Längsschnittstudie liegt zum Beispiel vor, wenn man eine Gruppe von Arbeitern mit einer bestimmten Schadstoffexposition über einen - vielleicht auch sehr langen - Zeitraum beobachten möchte. Bei einer Querschnittstudie würde jeder Arbeiter nur genau einmal befragt bzw. untersucht. Beide Studienarten tragen oft epidemiologischen Charakter:

Epidemiologische Studien befassen sich in erster Linie mit der Untersuchung von Krankheitsursachen bzw. mit der Wirkung von Risikofaktoren (vgl. auch Abschnitt 0.3). Typische Strukturen epidemiologischer Studien sind *Kohorten-* und *Fall-Kontroll-Studien*. Kohorten-Studien gehen von gesunden Personen aus, die unter einem Erkrankungsrisiko stehen (z.B. Arbeiter in der chemischen Industrie) und die über einen längeren Zeitraum beobachtet werden. Da das Erkrankungsrisiko in solchen Studien im Allgemeinen "relativ klein" ist, wird man deshalb eine im Allgemeinen

"relativ große" Kohorte beobachten müssen. Ein wesentliches Problem bei Kohorten-Studien sind *Studienaussteiger* (sog. "*dropouts*"), die sich – beispielsweise wegen Umzugs – einer weiteren Beobachtung entziehen: Niemand weiß, ob sich unter diesen Abbrechern vielleicht eine größere Anzahl erkrankter Personen befindet (der genannte Umzug kann krankheitsbedingt sein!) und damit eine *systematische Verzerrung* (engl. "*bias*") der Untersuchungsergebnisse zu befürchten ist. Fall-Kontroll-Studien beschäftigen sich primär mit der Ätiologie (den Ursachen) von Erkrankungen und umfassen eine Fall- und eine Kontrollgruppe, wobei jeder Person aus der einen Gruppe eine zweite Person aus der anderen Gruppe zugeordnet werden kann, die ihr in Bezug auf möglichst viele Eigenschaften möglichst ähnlich ist. Diese Paarbildung wird als *Matching* oder auch als *Matched-Pairs-Technik* bezeichnet. Auch bei Fall-Kontroll-Studien benötigt man im Allgemeinen recht erhebliche Kollektivumfänge, um statistisch fundierte Aussagen treffen zu können. Kohortenstudien sind prinzipiell prospektiver, Fall-Kontrollstudien dagegen retrospektiver Natur.

Wesentlich aufwendiger, aber auch in der Regel viel aufschlussreicher als Datenerhebungen sind *Experimente* bzw. *Versuche*. In einem Experiment, das grundsätzlich nur prospektiven Charakter besitzen kann, kann man *Einflussgrößen*, die die eigentlichen *Zielgrößen* (Merkmale) beeinflussen können, in kontrollierter Weise beliebig variieren. Zum Beispiel kann ein Untersucher als Zielgrößen die Herzfrequenz, den systolischen und den diastolischen Blutdruck vor und nach Medikation bestimmen und kann dabei als Einflussgrößen zum Beispiel unterschiedliche körperliche Belastungen auf dem Ergometer variieren, wobei etwa das Geschlecht und das Alter der Patienten bzw. Probanden als Kovariable dokumentiert wird.

Eine *Kontrollierte Klinische Studie* (engl. *Controlled Clinical Trial*, *CCT* oder *Randomized Clinical Trial, RCT*) ist eine prospektive Studie mit den wesentlichen Charakteristika eines Experiments: Ein typisches Beispiel für eine solche Studie ist der Vergleich zweier Therapien, der mit Hilfe von zwei Patientenkollektiven in zwei Studienarmen unter genau festgelegten Bedingungen gemäß eines verbindlichen *Studienprotokolls* durchgeführt wird. Über Details zur Planung und Auswertung von klinischen Studien wird im nächsten Kapitel gesprochen; dazu auch Kabisch et al. (2011).

Alle - wie auch immer - gewonnenen Daten sammelt man in der sogenannten *Urliste*. Eine Urliste umfasst alle *Basisdaten*, die sich im Verlaufe eines Experimentes, einer klinischen Studie oder einer Datenerhebung ergeben. Das bedeutet im obigen Beispiel insbesondere, dass ein Untersucher nicht nur zum Beispiel die Blutdruck*veränderung* erfasst, sondern auch explizit den Wert vor *und* den Wert nach einer körperlichen Belastung oder einer Medikation. Nur so kann man viele Fehler bei der Dokumentation vermeiden und gleichzeitig sicherstellen, dass einerseits die Daten später in überprüfbarer und optimaler Weise analysiert werden können und andererseits vielleicht auch einen Zugang zu Fragestellungen weiterer, detaillierterer Untersuchungen eröffnen.

1.3 Aspekte der Studienplanung und -auswertung

Nach verbindlicher Formulierung der Fragestellung einer Studie beschäftigt man sich als Erstes mit der statistischen *Planung* der Studie, wozu einige wesentliche Gesichtspunkte bereits in den Abschnitten 1.1 und 1.2 angesprochen wurden. Die konkreten Anforderungen der wissenschaftlichen Praxis werden ausführlicher in Kapitel 2 behandelt, die Kapitel 3 bis 6 sind der statistischen Auswertung von Studien gewidmet. Zum ersten Einstieg ist sicher ein kurzer Überblick nützlich:

Selbstredend sollte eine Datenerhebung oder ein Experiment nicht "einfach so" begonnen werden, sondern man wird sich in der Vorphase eingehend und sehr sorgfältig Gedanken über die Struktur und den Ablauf der Untersuchung machen. Natürlich ist ein solcher Entwurf in erster Linie durch die spezifische biologische bzw. medizinische Fragestellung der Untersuchung bestimmt, trotzdem muss man aber auch einige statistische Prinzipien bereits in der Planungsphase berücksichtigen, um auf diese Weise eine korrekte Struktur und eine spätere korrekte Auswertung der Untersuchung zu gewährleisten.

Eine erste grundlegende Überlegung beschäftigt sich stets mit der Frage, ob es vielleicht "äußere" Faktoren gibt, die Einfluss auf die Untersuchungsergebnisse besitzen können. Beispielsweise kann dabei der Zeitpunkt oder die zeitliche Länge der Untersuchung eine Rolle spielen, oder es ist denkbar, dass die Standards verschiedener Labors berücksichtigt werden müssen, ebenfalls kann es "innerhalb" eines Labors sinnvoll sein, die Tageszeit oder die mit den Analysen betrauten MTAs als "Störfaktoren" zu berücksichtigen. Werden in einer Studie Patienten untersucht, so wird man zweckmäßigerweise mögliche Einflussfaktoren wie Geschlecht, Alter, Anamnese und Begleiterkrankungen etc. berücksichtigen müssen. Die Statistik spricht hierbei von einer *Blockbildung*, deren Prinzip in Kapitel 2 eingehender beschrieben wird.

Als weitere grundsätzliche Regel gilt in der Studienplanung die Anwendung des *Zufallsprinzips*. Dieses wendet man bei der Auswahl von Studien-Patienten oder Probanden an, andererseits spielt es auch bei der Zuteilung von Behandlungen o.ä. eine wichtige Rolle. Dieses Prinzip der *Randomisierung* ist Gegenstand von Abschnitt 2.4.

Nicht zuletzt muss man bei der Planung einer Untersuchung auch an die späteren Auswertungsmethoden denken: Die sachgerechten Methoden zur Datenauswertung kann man nur finden, wenn man alle relevanten Daten, wie bereits im Abschnitt 1.1 ("Statistische Skalen und Daten") beschrieben, detailliert klassifiziert. Damit eng verbunden ist die korrekte Planung des erforderlichen Stichprobenumfanges ("*Fallzahlberechnung*"), die speziell bei klinischen Studien eine zentrale Rolle spielt. Einzelheiten dazu finden sich in den Abschnitten 2.3, 4.5 und 5.10, die sich mit den für eine Studie adäquaten Fallzahlen auseinandersetzen.

Ein erster Auswertungsschritt - obligatorisch für jede Studie - ist die sogenannte *Deskriptive Datenanalyse*. Diese umfasst das Anlegen von Tabellen, das Darstellen von Histogrammen, Box-Plots, Kreisdiagrammen oder von Zeitverlaufsreihen und Anderes mehr. Solche Darstellungen sind einfach mit Hilfe von einschlägigen Computerprogrammen möglich: Viele Anwender haben sich dabei auf das Programm **Excel** kapriziert, das nicht nur als Tabellenkalkulationsprogramm gute Dienste leistet, sondern auch brauchbare Graphiken anbietet. Leider hat sich bei verschiedenen Herstellern manche eigenwillige Darstellungsweise etabliert, so dass nicht alles Angebotene unbedingt empfohlen werden kann: Darauf wird in diesem Text vielfach hingewiesen. Die speziellen Techniken einer korrekten Analyse werden in den nächsten Abschnitten ausführlich dargestellt.

Das weitaus größere Gebiet der Statistik ist die *Konfirmatorische Statistik*, gelegentlich auch *Induktive Statistik* oder auch ganz einfach *Teststatistik* genannt. Dieses Gebiet beschäftigt sich mit den sogenannten *statistischen Testverfahren*, mit deren Hilfe man konkrete Arbeitshypothesen, sogenannte *Nullhypothesen*, überprüfen kann. Den ersten Satz des Vorwortes "Das kann kein Zufall sein!" würde ein Biometriker sinngemäß umsetzen in die Nullhypothese (bitte beachten Sie die Formulierung!) "Das ist ein Zufall" und würde im Anschluss auf Basis eines geeigneten statistischen Tests versuchen, diese Nullhypothese zu stützen oder zu widerlegen: Ein solches Testverfahren vermittelt eine objektive Entscheidungsgrundlage, ob man einen gegebenen Befund beziehungsweise eine vorliegende Beobachtung noch durch den *Zufall* oder schon durch eine vielleicht vorhandene Gesetzmäßigkeit erklären sollte. Im Beispiel der Antihypertensivumstudie in Abschnitt 1.2 wird der Untersucher die statistische Nullhypothese "Keine Blutdruckveränderung nach Medikation" formulieren, sich dazu einen geeigneten statistischen Test auswählen (mehr dazu in Kapitel 5) und anhand dieses objektiven Testverfahrens zu einer Entscheidung kommen, ob er die beobachteten Blutdruckveränderungen als wesentlich im Sinne einer Medikamentenwirkung oder eher als unwesentlich, vielleicht noch durch die biologische Variabilität o.ä. erklärbar auffassen sollte.

Der offensichtliche Vorteil einer teststatistischen gegenüber einer nur deskriptiven Analyse besteht darin, dass alle Entscheidungen von subjektiven Einflüssen unbeeinflusst sind und insbesondere jeder Untersucher und auch jeder spätere Leser der Studie zu dem gleichen Urteil gelangt bzw. die Studie in gleicher Weise interpretieren kann. Einige wichtige Testverfahren werden exemplarisch in den Kapiteln 5 und 6 beschrieben. Um einen Einblick in die Konstruktionsweise solcher Tests zu erhalten, wird dort für die drei relevanten Skalentypen (das sind nominale, ordinale und quantitative Skalen) jeweils ein Vertreter etwas ausführlicher dargestellt. In diesem Zusammenhang wird auch der in der Statistik zentrale Begriff "Nullhypothese" formal definiert: In diesem formalen Rahmen werden die eben so genannten "wesentlichen" Unterschiede, wie in der Statistik üblich, als "*statistisch signifikante Unterschiede*" bezeichnet werden.

Wie bei deskriptiven Methoden ist es natürlich erst recht für statistische Testverfahren wünschenswert, die Auswertung nicht gerade mit Papier und Bleistift vornehmen zu müssen, wozu zahlreiche, zum Teil auch sehr umfangreiche statistische Programmpakete in unterschiedlichen Preiskategorien verfügbar sind. Bekannte "große" Vertreter sind zum Beispiel **SAS** oder **SPSS**, unter den kleineren finden sich auch interessante "AddOns" für **Excel** – zum Beispiel **WinStat** - und natürlich viele andere, mehr oder weniger umfangreiche und mehr oder weniger kostenträchtige Alternativen. Der Autor dieses Skriptums hat seinerseits ein Programmpaket entwickelt ("**BiAS.**", d.h. Biometrische Analyse von Stichproben), dies in der Absicht, möglichst alle Methoden, die man mehr oder weniger täglich benötigt, in einem möglichst einfach zu handhabenden Programm zusammenzufassen. Das Programm **BiAS.** umfasst deskriptive Methoden, viele Testverfahren und viele Verfahren zur statistischen Studienplanung einschließlich Power-Berechnungen und der Planung von Stichprobenumfängen ("Fallzahlberechnungen"). Alle Abbildungen in diesem Skriptum wurden - von wenigen Ausnahmen abgesehen - mit **BiAS.** erzeugt, ebenfalls wurden sämtliche Berechnungen zu allen Beispielen mit **BiAS.** durchgeführt.

Abschluss jeder Datenerhebung und jeder klinischen oder epidemiologischen Studie ist der *Abschlussbericht*, der nicht nur die Ergebnisse, sondern auch alle Einzelheiten zur statistischen Versuchsplanung und zur Durchführung der Studie enthält. Zur Form eines solchen Berichtes spielen sicher dienstliche und vielleicht auch persönliche Gepflogenheiten eine Rolle, so dass hier auf dieses Thema nicht eingegangen wird. Formale Regeln zur Durchführung von klinischen Studien beschreiben die ICH- bzw. die GCP-Richtlinien (ICH = *International Commitee of Harmonisation*, GCP = *Good Clinical Practice*, ergänzt durch GSP/GMP = *Good Statistical/ Manufacturing Practice*).

1.4 Entwicklungsphasen medizinischer Studien

Die statistische Planung medizinischer und pharmakologischer Studien hängt davon ab, in welcher Phase der Entwicklung zum Beispiel eines Medikamentes man sich befindet. Dies veranschaulicht das üblicherweise verwendete Modell der vier Entwicklungsphasen eines Pharmakons, das ohne Weiteres auch auf andere medizinische Fragestellungen übertragen werden kann:

Phase I: Diese Versuche finden im Bereich der Klinischen Pharmakologie und Toxikologie statt und beschäftigen sich primär mit der Arzneimittelsicherheit, weniger mit der Effizienz. Die Versuche werden gewöhnlich mit Freiwilligen (d.h. mit *Probanden*) durchgeführt und dienen unter anderem der Dosisfindung unter Berücksichtigung von Nebenwirkungen. Nicht zuletzt wird der Metabolismus und die Bioverfügbarkeit der pharmakolo-

gischen Substanzen untersucht (Anhang, AUC!). In Phase I trifft man sehr häufig Versuchspläne mit mehreren Behandlungen pro Proband/Patient an (*Cross-Over-Versuche*, eine präzisere Definition folgt in Abschnitt 2.2), da, wie erwähnt, in Phase I mit freiwilligen Probanden gearbeitet wird und solche Versuchsstrukturen deshalb aus sachlichen, aber auch aus Kostengründen adäquat sind. Eine Randomisierung ist verbindlich, eine Fallzahlberechnung zwar erstrebenswert, wegen fehlendem Vorwissen (vgl. auch die Abschnitte 4.5 und 5.10!) aber nicht immer möglich.

Phase II: Diese Versuche werden mit ausgewählten Patienten durchgeführt und ähneln ansonsten den Studien der Phase I. Stets wird ein sorgfältiges *Monitoring* der Patienten durchgeführt, um Aufschlüsse über die Effektivität und Sicherheit der Pharmaka zu gewinnen. Eine Randomisierung ist auch hier wie Blockbildung verbindlich, eine Fallzahl- bzw. Powerberechnung ist erstrebenswert.

Phase III: Nachdem in Phase II die Wirksamkeit eines Medikaments bestätigt wurde, finden in Phase III *Kontrollierte Klinische Studien* zur Sicherung dieser Erkenntnis und zum Vergleich des neuen Pharmakons mit konkurrierenden Standards statt. Der Vergleich mit *Placebo* (also einem pharmakologisch wirkungslosen "Leerpräparat") ist in Phase III nicht nur aus ethischen Gründen obsolet und findet eher in Phase I oder II eine Berechtigung. Fallzahlberechnung und Randomisierung sind wie alle anderen Prinzipien der statistischen Versuchsplanung verbindlich.

Phase IV: Nachdem ein neues Medikament eine Zulassung durch die entsprechende Behörde erhalten hat (in Deutschland: das Bundesinstitut für Arzneimittel und Medizinprodukte BfArM als Nachfolgeorganisation des BGA, in den USA das FDA), werden Phase-IV-Studien unternommen. In dieser Phase werden in *Feldstudien* sehr große Kollektive untersucht, um das nunmehr zugelassene Medikament auf seltene Nebenwirkungen und bezüglich Morbidität und Mortalität (vgl. Abschnitt 3.4) hin zu prüfen. Phase-IV-Studien tragen fast stets epidemiologischen Charakter, die Versuchsplanung ist häufig problematisch.

Vor den beschriebenen Humanversuchen werden in einem vorklinischen Stadium in der Regel Tierversuche zur Beurteilung des Pharmakometabolismus, der potentiellen Wirksamkeit und der möglichen Toxizität durchgeführt. Statistische Prinzipien wie die der Blockbildung und Randomisierung müssen auch hier unbedingt beachtet werden, Fallzahl- bzw. Powerberechnungen sind ebenfalls verbindlich.

Die verschiedenen Aspekte der statistische Versuchsplanung sind in allen Entwicklungsphasen grundsätzlich die gleichen, werden aber unter Umständen von den speziellen Fragestellungen bestimmt. Die grundlegenden statistischen Prinzipien *repräsentative Stichproben*, *Randomisierung* und *Blockbildung* gelten in allen Bereichen!

Kapitel 2:
Studienplanung

Die statistische Studienplanung beschäftigt sich mit der Strukturierung von Datenerhebungen, von epidemiologischen und von klinischen Studien, andererseits mit Methoden der Stichprobennahme und nicht zuletzt mit der Randomisierung von Versuchsbedingungen (z.B. von Therapien). Der so skizzierte "Königsweg" der *Versuchsplanung* ist vollumfänglich leider nur kontrollierten Studien erschlossen und in aller Regel bei Datenerhebungen nur eingeschränkt gangbar. Vielfach ist aber sogar bei klinischen Studien bereits die Zielvorstellung repräsentativer Stichproben problematisch, und so erfordert die statistische Planung und Analyse nicht nur von Datenerhebungen, sondern auch von klinischen und epidemiologischen Studien eine gewisse Sensibilität für die theoretischen Voraussetzungen aller Methoden. Zum Verständnis sind noch einige Sachverhalte zu klären:

2.1 Stichprobe und Grundgesamtheit

Ein Experiment oder eine Klinische Studie ist "für sich betrachtet" nur von minderem Interesse, vielmehr sind solche Untersuchungen nur deshalb relevant, weil man unter Umständen aus den Resultaten über den Versuch hinausgehende Schlüsse ziehen kann. Lässt man einen Gegenstand mehrfach fallen und er fällt immer zu Boden, so wird man vielleicht den Schluss ziehen können, dass dies "immer" der Fall sein wird, vielleicht sogar mit beliebigen Gegenständen. Beobachtet man an einer Gruppe von Patienten einer definierten Diagnose gewisse Eigentümlichkeiten, so ist man versucht zu sagen, dass dies vielleicht auch bei allen anderen Patienten dieser Diagnose der Fall ist: Zur Objektivierung definiert man eine *Arbeitshypothese* und versucht, diese auf Grundlage der empirischen Ergebnisse zu stützen oder zu widerlegen. Im ersten Beispiel fällt dies sicher nicht schwer, wie und wann aber kann man einen verallgemeinernden Schluss in medizinischen oder pharmakologischen Studien ziehen? Zur Klärung des "wann" ist zunächst eine Modellbetrachtung erforderlich, das "wie" wird sich später in Kap. 5 im Rahmen des statistischen Testens aufklären.

Wie groß sind neugeborene Kinder im Allgemeinen? Welcher Therapieerfolg (z.B. welche Blutdrucksenkung) ist auf Grund einer Behandlung zu erwarten? Welche Laborwerte sind künftig als "normal" zu bezeichnen? Solche und ähnliche Fragestellungen können Anlass für eine wissenschaftliche Studie sein. Dabei steht aber offenbar nicht so sehr die Frage im Vordergrund, wie die Untersuchungsergebnisse speziell bei Meier und Müller ausfielen, sondern man interessiert sich vielmehr dafür, wie sich ein Phänomen "im Allgemeinen" darstellt, wie sich also eine viel größere

Gruppe von gleichartigen bzw. "ähnlichen" Personen verhalten mag. Aufschluss darüber möchte man mit Hilfe einer Stichprobe erhalten.

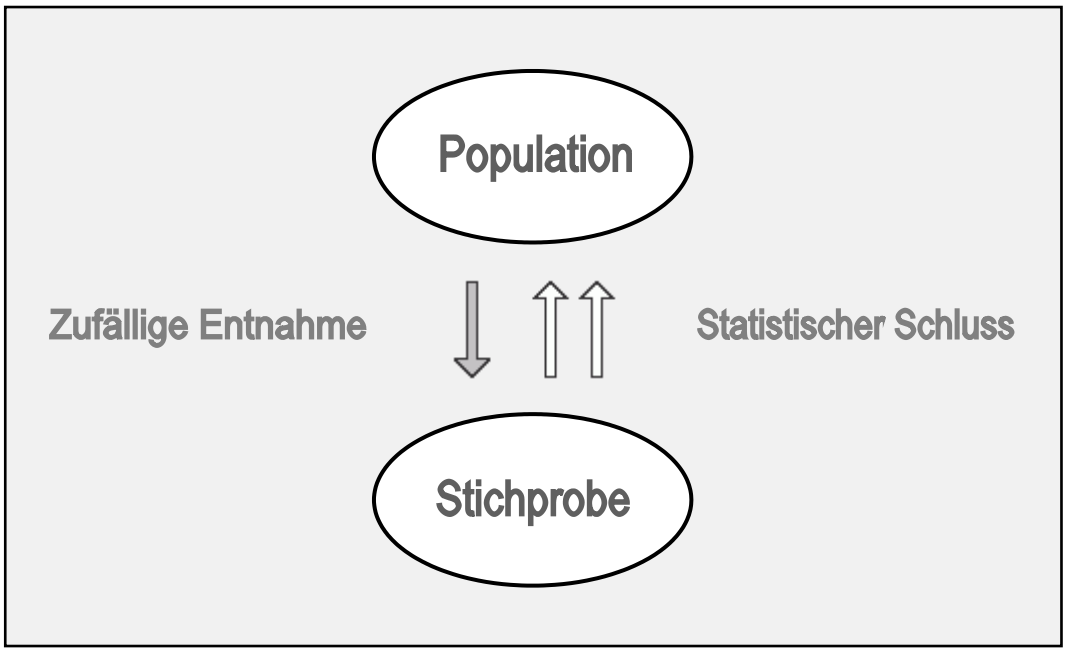


Abbildung 4: Modell „Population und Zufallsstichprobe“

Dem theoretischen Modell folgend, geht man stets von einer definierten *Grundgesamtheit* (synonym: *Population*) aus und zieht aus dieser Grundgesamtheit eine repräsentative *Stichprobe*. Die Population wird man sehr sorgfältig definieren, denn über diese möchte man ja eine Aussage treffen. Im Beispiel des Antihypertensivums könnte die Population aus allen denkbaren Patienten einer bestimmten Altersgruppe und eines definierten Schweregrades ohne bekannte Sekundärdiagnosen bestehen; diese Einschränkungen können sinnvoll sein, wenn man den Einfluss möglicher *Störfaktoren* ausschalten möchte (mehr dazu in den Abschnitten 6.10-13). Aus der definierten Population zieht man nun - im Idealfall mit Hilfe eines Zufallsprozesses - eine Stichprobe. Diese zufällig, also unsystematisch gewonnene, repräsentative Stichprobe (nicht-repräsentative Stichproben sind prinzipiell sinnlos) gibt ganz sicher wenigstens angenähert die Verhältnisse in der Grundgesamtheit wieder und kann in gewisser Weise als "Modell" für diese Grundgesamtheit aufgefasst werden. Damit ist es auch plausibel, dass man aus den konkreten Stichprobenergebnissen auf die Verhältnisse in der (unbekannten!) Grundgesamtheit schließen darf: Beobachtet man in der Stichprobe einen gewissen Behandlungseffekt, so schließt man daraus vielleicht, dass auch bei den zukünftigen Patienten obiger Definition "im Allgemeinen" ein Heilungserfolg eintreten wird - um diesen Schluss auch "objektiv" durchzuführen, wird weiter unten in Kapitel 5 ein statistisches Testverfahren verwendet.

In der Praxis stellt sich die Situation in der Regel etwas schwieriger dar, denn wie sollte man auf die Patienten der oben definierten Population zugreifen können, um daraus eine repräsentative Stichprobe zu ziehen? Allein die Tatsache, dass es im Beispiel auch latent erkrankte Personen gibt, die von ihrer Erkrankung noch gar nicht wissen, macht dies unmöglich, selbst wenn man davon absieht, dass man ganz sicher nicht alle sich ihrer Erkrankung bewussten Patienten erreichen könnte und, noch weniger, alle Patienten prinzipiell zu einer Teilnahme an der Untersuchung bewegen könnte. Also ist man auf Patienten angewiesen, die man in einem Krankenhaus oder in einer Praxis vorfindet und muss sich in Umkehrung des Modells überlegen, für welche Population die vorgefundenen Patienten "repräsentativ" sein könnten. Dabei sind möglichst alle studienrelevanten Einflussgrößen (Alter, Geschlecht, Begleiterkrankungen, Alkoholkonsum, Rauchen,…) zu berücksichtigen, um das erstrebte Modell noch als plausibel zu akzeptieren. Selbstredend kann man dabei - vielleicht auch aus Unkenntnis - wichtige Einflussgrößen übersehen, womit sich eine grundlegende Schwäche dieses Vorgehens zeigt. Trotzdem ist dieser Weg der einzige, der in praxi eingeschlagen werden kann.

Zusammenfassend definiert man also als *Zielgruppe* einer Studie die Population aller denkbaren Patienten bzw. Probanden mit umschriebenen Eigenschaften, betrachtet eine repräsentative Stichprobe, studiert das Verhalten bzw. die medizinischen Eigentümlichkeiten dieser Stichprobe und zieht vermöge eines statistischen Verfahrens verallgemeinernd Rückschlüsse auf die Population.

In einer Studie kann man auch eine ganze Anzahl von *Subpopulationen* (*Strata, Singular: Stratum*) gleichzeitig berücksichtigen. Die korrespondierenden Teil-Stichproben bezeichnet man als *stratifizierte Stichproben* und spricht dabei auch von einer *Schichtung*. (Der an dieser Stelle gelegentlich verwendete Begriff *Blockbildung* bezieht sich eher auf den Kontext der randomisierten Blöcke in Abschnitt 2.4 und sollte hier vermieden werden.) Die oben beschriebenen Zusammenhänge sind prinzipiell auch auf geschichtete Stichproben zu übertragen, wenn auch die spätere Analyse solcher Studien etwas aufwendiger ist.

2.2 Studiendesign

Zur Festlegung eines Studiendesigns gehören neben den letzten Überlegungen auch eine Festlegung von weiteren Strukturen wie die zeitliche Abfolge der Studie, ob man die Studie unter Umständen vorzeitig abbrechen muss (es könnte z.B. sein, dass unerwarteter Weise schwere Nebenwirkungen einer Behandlung auftreten!), die *Blockbildung* (im Therapievergleich möchte man innerhalb eines gewissen Zeitraumes gleich viele Patienten mit Behandlung A bzw. B behandeln oder A und B

den "Schichten" Frauen und Männern gleich oft zuteilen) und anderes mehr. Speziell sind auch Überlegungen erforderlich, wie man die Behandlungen oder Versuchsbedingungen den Patienten zuteilen sollte ("*Randomisierung*", nächster Abschnitt). Eine nicht unwesentliche Rolle spielt die Überlegung, ob man einen *Parallelgruppenversuch* oder aber eine sogenannte *Cross-Over-Studie* durchführen möchte:

In einem *Parallelgruppenversuch* vergleicht man zwei oder mehrere Therapien mit Hilfe von ebenso vielen Patientengruppen (Stichproben), wobei jeder Patient in nur genau einer Stichprobe enthalten ist; verschiedene Therapien am gleichen Patienten sind damit ausgeschlossen. Studien dieses Typs werden gelegentlich auch als *zweiarmige* bzw. als *mehrarmige Studien* bezeichnet.

Eine Untersuchung von nur einer Stichprobe (also der Verzicht auf eine *Kontroll-* bzw. *Vergleichsgruppe*) ist nicht empfehlenswert. Im Beispiel der Blutdrucktherapie ist zwar zunächst nur der Vergleich vor und nach Therapie interessant, trotzdem kann man nicht sicher sein, ob während der Behandlungsphase noch andere Einflussfaktoren - ganz elementar vielleicht die Wetter- bzw. Luftdrucklage o.ä. - eine Rolle spielen und die Untersuchungsergebnisse davon - und womöglich weniger von den beiden Therapien - beeinflusst sind. Um die Untersuchungsergebnisse vom *Einfluss äußerer Wirkungen* zu bereinigen, sollte man deshalb stets eine Kontrollgruppe in der Studie mitführen. Damit ergibt sich gleichzeitig ein Vergleich mit einer Standardtherapie: Ziel ist somit die Untersuchung zweier strukturell möglichst identischer Studiengruppen ("*Studienarme*"), die sich nur durch die unterschiedlichen Therapien unterscheiden.

In einer *Cross-Over-Studie* wird jeder Patient mit (nach Möglichkeit:) allen versuchsrelevanten Therapien behandelt, wobei - im Falle zweier Therapien, bei mehreren Therapien gilt das gleiche - die *Behandlungssequenzen* AB und BA gleich oft vorkommen sollten. Wie im Parallelgruppenversuch ist es auch hier angeraten, sog. *Zeitblöcke* zu bilden, denn nur so kann man sicherstellen, dass in einem Zeitraum alle Therapien bzw. Therapiesequenzen gleich oft vorkommen und damit ein möglicher Einfluss des Zeitfaktors ausgeschlossen ist. Darauf wird im nächsten Abschnitt wieder zurückgekommen.

Cross-Over-Studien sind mitunter problematisch, denn möglicherweise besitzt die erste Behandlung einen Einfluss auf die zweite, und unter Umständen hängt dies zusätzlich davon ab, ob zuerst "A" oder "B" verabreicht wurde. Solche *Übertragungseffekte* (synonym: *Carry-Over-Effekte*) sind nicht einfach zu behandeln und nach Möglichkeit auszuschließen. Dies kann man einerseits statistisch untersuchen, andererseits, was sicher die bessere Lösung darstellt, unter medizinischen Kriterien ausschließen. Cross-Over-Designs sollten somit möglichst nur mit gesunden Probanden oder mit Patienten im *Steady-State* (bei denen sich eine Erkrankung im Laufe der Zeit nicht verändert) durchgeführt werden, außerdem sollte man eine ausreichende *Wash-Out-Periode* (Eliminations-

phase) einhalten, innerhalb derer die Wirkung der ersten Behandlung sicher verschwindet (Pause nach Elimination der Wirksubstanz!). Allerdings sind die möglichen Einsatzgebiete damit eher rar und vorwiegend nur in Phase-I-Studien (vgl. Abschnitt 1.4) und/oder bei speziellen Erkrankungen wie zum Beispiel Asthma Bronchiale denkbar, deren Schweregrad im Wesentlichen gleich bleiben mag: Die jeweils zweite Behandlung trifft idealerweise den gleichen Patienten im gleichen Zustand an wie die jeweils erste Behandlung.

Im Zweifelsfall stellt die Planung eines Parallelgruppenversuchs die bessere Lösung dar. In einem Cross-Over-Versuch erhält man pro Proband/Patient natürlich mehr Daten als im Parallelgruppenversuch, womit man offenbar Versuchspersonen "sparen" kann. In aller Regel muss man aber in der Cross-Over-Analyse sicherstellen, dass tatsächlich keine Übertragungseffekte (z.B. Nachwirkungen der Erstbehandlung) vorhanden sind und benötigt zu dieser Prüfung unter Umständen mehr Patienten als für einen Parallelgruppenversuch, worauf in diesem Skriptum allerdings nicht weiter eingegangen wird. Als letzte "Rettung" einer Cross-Over-Studie, in der sich vielleicht Übertragungseffekte herausstellen, kann man die erste Phase (i.e., Erstbehandlung) wie in einem Parallelgruppenversuch auswerten und die zweite Phase (Zweitbehandlung) ignorieren, was aber sicher nicht im Sinne einer ökonomischen Versuchsdurchführung ist und eher als "Kunstfehler" einzustufen ist.

Auf Cross-Over-Designs wird im Weiteren nicht mehr Bezug genommen, bei Interesse sei dazu auf die einschlägige Literatur verwiesen. Dort finden sich auch Hinweise auf Drei-Phasen-Cross-Over-Designs (ABB und BAA), die einige der genannten Schwächen nicht besitzen.

Bei vielen Fragestellungen macht man sich das Prinzip des *Matchings* zu Nutze, denn vielfach sind Untersuchungen an Zweierblöcken effizienter als Studien an unabhängigen Studiengruppen. Ein solcher Block kann in einer dermatologischen Studie der linke und rechte Arm des Patienten sein, man kann am gleichen Patienten den Blutdruck vor und nach Therapie messen, und so gibt es noch viele andere Möglichkeiten (z.B. Geschwister- oder Alterspaare), sogenannte *verbundene Stichproben* herzustellen. Ein spezielles Auswertungsverfahren wird in Abschnitt 5.3 diskutiert.

Um alle subjektiven Einflüsse auszuschließen, wird man neben einer "Randomisierung" (Abschnitt 2.4) das Studiendesign nach Möglichkeit "*blind*" oder optimal "*doppelt-blind*" strukturieren. Ein "blindes" Design bedeutet, dass die teilnehmenden Patienten nicht wissen, welches Medikament sie erhalten, "doppelt-blind" bedeutet, dass weder die Patienten noch der behandelnde Arzt wissen, welche Behandlung verabreicht wurde. Erst nach Versuchsende wird zur Auswertung die Zuteilung bekannt gegeben ("geöffnet"). Natürlich ist – zum Beispiel bei physikalischen Behandlungen - eine solche Konstruktion mitunter nicht möglich. Man spricht dann auch von sogenannten *offenen Designs*.

2.3 Fallzahlberechnung

Eine wesentliche Aufgabe bei der Planung einer Studie ist die Festlegung des erforderlichen Stichprobenumfanges, die nicht willkürlich, sondern immer auf Grundlage einer adäquaten *Fallzahlberechnung* erfolgen muss: Detallierte Angaben dazu werden von jeder Ethik-Kommission, dem BfArM und nicht zuletzt von wissenschaftlichen Journalen zwingend verlangt. Die Grundgedanken einer Fallzahlberechnung werden in den Abschnitten 4.5 und 5.10 exemplarisch für Konfidenzintervalle und t-Tests (und für weitere Verfahren ohne Herleitung) dargestellt. Grundsätzlich existiert zu jedem statistischen Testverfahren auch eine Methode zur Fallzahlberechnung.

Warum muss man eine Fallzahlberechnung durchführen? Es ist sicher einsichtig, dass man zum Beispiel in einem Vergleich zweier Schmerztherapien einen Wirkungsunterschied nicht ohne Weiteres mit vielleicht 5+5 Patienten "nachweisen" kann, im Gegenteil wird man sich dazu intuitiv eine möglicherweise deutlich größere "Fallzahl" vorstellen. Wünschenswert ist jedoch eine objektive, sachorientierte Festlegung der tatsächlich erforderlichen Fallzahl: Eine Fallzahlberechnung kann, nach Vorgabe des medizinisch relevanten Wirkungsunterschiedes zwischen den Therapien, im Sinne einer ökonomischen Versuchsdurchführung nur mit einschlägigen mathematischen Methoden der statistischen Stichprobenumfangsplanung vorgenommen werden. So gesehen gibt es keine "kleinen" oder "große" Fallzahlen, sondern, statistisch korrekt geplant, nur "richtige" Fallzahlen.

Häufig bleibt die Stichprobenumfangsplanung auch eine mathematische und medizinische Illusion, da in der Praxis oft noch viele limitierende Faktoren im Spiel sind. Elementare Faktoren sind etwa die Zeitdauer einer Studie oder auch der Kostenfaktor: Bekanntlich sind Laboranalysen mitunter recht kostenintensiv, während für eine konkrete Studie nicht nur im Rahmen des Routinebetriebes, sondern auch über sogenannte "Drittmittel" nur begrenzte Resourcen verfügbar sind. Für manche Studien stehen auch nicht so viele Patienten zur Verfügung, wie nach einer statistischen Fallzahlberechnung an sich erforderlich sein müssten. Dies soll aber nicht bedeuten, dass eine Fallzahlbestimmung mehr oder weniger überflüssig ist, denn immerhin dient diese der Gewährleistung, dass man bestimmte, "nachzuweisende" Effekte wie zum Beispiel die *minimale, medizinisch relevante Wirkung* eines Medikaments auch als "statistisch signifikant" darstellen kann. Diese Gedanken werden, wie bereits eingangs erwähnt, speziell in Abschnitt 5.10 wieder aufgegriffen.

2.4 Randomisierung

Der Begriff "*Randomisierung*" ist der englischen Sprache entlehnt und leitet sich aus dem Wort "random" ab, das bekanntlich soviel wie "zufällig" bedeutet. Neben der "zufälligen" Stichprobennahme, die zu repräsenta-

tiven Stichproben führt, spielt "der Zufall" auch in der Versuchsplanung bei der Zuteilung von Behandlungen o.ä. auf die Merkmalsträger (Patienten, Probanden, Versuchstiere etc.) eine Rolle.

Unter einer Randomisierung versteht man eine Zuteilung von Behandlungen bzw. "Bedingungen" auf Merkmalsträger (Patienten etc.), die frei ist von subjektiven Einflüssen und die ohne jede Systematik mit Hilfe eines Zufallsprozesses durchgeführt wird.

Weshalb ist dieses Vorgehen erforderlich?

Ein Pharmakologe plant eine Analgesiestudie bei Neuralgie-Patienten und möchte zwei Wirksubstanzen vergleichen, über deren möglicherweise unterschiedliche Wirkung er noch im Zweifel ist. Er wählt als Versuchsdesign einen Parallelgruppenversuch, legt die Stichprobenumfänge n_1 und n_2 für die beiden Studienarme fest (vgl. dazu auch Abschnitt 2.3 zur Fallzahlberechnung!) und teilt die beiden Substanzen, hier wieder mit "A" und "B" bezeichnet, "irgendwie" zu, dies möglicherweise mit dem zweifelhaften Erfolg, dass er - bewusst oder auch unbewusst - die Zuteilung von seiner subjektiven Einschätzung der Behandlungsbedürftigkeit der Patienten abhängig macht: Die Untersuchungsergebnisse bzw. die Behandlungseffekte würden auf diese Weise von den (beiden!) subjektiven Vorstellungen des Arztes beeinflusst (in der Biometrie spricht man dabei von einer *Vermengung von Effekten*). Um diesen oder ähnliche Einflüsse auszuschalten, wählt man grundsätzlich eine zufällige Zuteilung mit Hilfe von *Zufallszahlen*. Warum aber keine systematische Zuteilung, die ebenfalls die Subjektivität des Arztes ausschalten könnte?

Als systematische Zuteilung käme vielleicht eine Verabreichung des Medikamentes "A" an die zuerst in die Studie aufgenommenen Patienten und entsprechend "B" für die zuletzt aufgenommenen in Frage. Möglicherweise kommt hierbei jedoch ein Zeittrend zum Tragen, der die Untersuchungsergebnisse verfälscht, ganz sicher aber spielt wieder ein Lerneffekt seitens des Arztes, eventuell auch seitens der Patienten eine Rolle. Immerhin ist klar, dass zuerst das eine, später das andere Medikament verabreicht wurde, und die beobachteten Schmerzäußerungen der Patienten können Anlass zu einer subjektiv oder objektiv veränderten Zuwendung seitens des Arztes führen, die wiederum die Patienten beeinflusst. Auch dies ist keine erstrebenswerte Situation, denn auch äußere Einflüsse möchte man aus den Studienresultaten fernhalten. Systematische Zuteilungen, gleichgültig ob in der beschriebenen oder in einer anderen Form, können also ebenfalls zu Verfälschungen der Untersuchungsergebnisse führen.

Die Randomisierung ist als eine Art "Versicherung" aufzufassen, für die man einen geringen Preis zahlt (nämlich die Verwendung von Zufallszahlentabellen oder von Programmpaketen mit einschlägigen Funktionen, zum Beispiel **BiAS**.), sich aber damit auf einfache Weise gegen "Pannen" wie die oben skizzierten absichern kann.

Im Beispiel der Analgesie-Studie bei Neuralgie-Patienten kann man sich den Prozess der randomisierten Zuteilung einfach verdeutlichen:

Der Leiter der Analgesiestudie entschließt sich aus medizinischen Gründen, alle Patienten, die innerhalb etwa einer Woche in die Studie aufgenommen werden, als einen Zeitblock aufzufassen. Ein solcher Block mag aus vier Patienten bestehen (andere Anzahlen, falls zweckmäßig, sind gleichermaßen denkbar), und so ist es die Aufgabe des Versuchsleiters, auf jeweils vier Patienten zweimal "A" und zweimal "B" zu randomisieren. Da er einen Doppeltblind-Versuch durchführt, wird die Randomisierung von einem Dritten vorgenommen, der mit Hilfe von *Zufallszahlen* die Zuteilung definiert. Zur Veranschaulichung kann man sich vorstellen, dass die Hilfskraft einen Würfel zur Erzeugung von Zufallszahlen verwendet und vereinbart, dass den Zahlen 1-3 die Therapie "A" und den Zahlen 4-6 Therapie "B" entspricht. Ist eine Therapie innerhalb eines Zeitblockes bereits zweimal vorgekommen, so erhält der letzte bzw. erhalten die beiden letzten Patienten des Blockes zwangsläufig die andere Therapie.

Zufallszahl	**1**	**2**	**3**	**4**	**5**	**6**
Therapie	**A**	**A**	**A**	**B**	**B**	**B**

Patient	**Zufallszahl**	**Therapie**
1	**6**	**B**
2	**1**	**A**
3	**4**	**B**
4	**2**	**A**
5	**3**	**A**
6	**5**	**B**

Tabelle 1: Beispiel einer Randomisierung

In Tabelle 1 wird eine Randomisierung auf zwei Gruppen mit den Stichprobenumfängen n_1=3 und n_2=3 durchgeführt; im Beispiel ist keine Blockbildung vorgesehen. In der ersten Teiltabelle wird vereinbart, dass den Zufallszahlen 1-3 Therapie A und 4-6 Therapie B entsprechen soll. Man erhält der Reihe nach die Zufallszahlen 6,1,4,2,3,5 und ordnet deshalb in der zweiten Teiltabelle den 6 Patienten der Reihe nach die Therapien B,A,B,A,A,B zu. Die Zufallszahlen wurden hier beispielhaft mit einem Würfel erzeugt und es wurde dabei eine bereits gewürfelte Zahl bei späteren Würfen ignoriert.

Da das Würfeln vielleicht nicht unbedingt als die Methode der Wahl aufzufassen ist, empfehlen sich für die Praxis *Zufallszahlentabellen* (zum Beispiel die Tabelle 2) oder, wie oben erwähnt, einschlägige Computerprogramme; die geschilderte Struktur der randomisierten Zuteilung bleibt die gleiche. In manchen Lehrbüchern finden sich umfangreiche Zufallszahlentabellen unterschiedlicher Struktur, auf die man gegebenenfalls

zurückgreifen kann, um nicht immer die gleichen Zahlen zu verwenden. Optimal ist sicher eine computerunterstützte Randomisierung.

Von einem Computer erzeugte Zufallszahlen werden häufig auch als *Pseudo-Zufallszahlen* bezeichnet, da sie zwar in aller Regel nach einem bestimmten Algorithmus errechnet werden, trotzdem aber alle Tests auf "Zufälligkeit" bestehen.

Tabelle 2 zeigt eine Tabelle mit Zufallszahlen, die zur Randomisierung verwendet werden kann. In jeder der 20 Zeilen stehen je 5 Blöcke zu jeweils 10 Zahlen, wobei in jedem Block jeweils alle Zahlen von 0 bis 9 vorkommen, und zwar genau einmal. Man spricht deshalb auch von einer *Zufallspermutation* der Zahlen von 0 bis 9.

9251804376	8364925107	9412857036	3427095618	8410973625
6154209837	7294385601	1298347560	5408972631	8462957013
6705493281	6417382095	3956247108	9260485713	4930718256
2413697805	6054829731	7614058923	7198240536	6723814590
2368940751	8945261703	8241530796	1486937025	6438207915
8365142079	2857946130	2840639175	9614083572	8473509621
0653427891	3209467815	1403596827	7860241395	3498750612
0759241368	5679320184	0216348975	2397546018	4089261375
4260359781	1370856249	3890561472	0682351947	4865912730
0874596231	4982761035	5724089631	4187960352	7631845209
6753924081	9260835417	4968721530	8476305219	6379814025
0578241396	6934718025	1745698320	8496520731	9184052736
0263518749	8691234705	8139524706	0281975436	3619702584
6510837294	3940268517	1928364057	5027986431	7315092684
4509278613	6594783102	7968154302	3024179658	0492316758
9384017256	3708125469	9183254076	1362950874	2763085941
4781659203	3457608192	4971260538	4031628957	4387620591
4259708613	5286791403	4317290568	8107632945	5164793028
6028945713	5280316479	0421597683	7039486215	9763524081
0914372568	6743219580	1652389470	8321960457	1879034256

Tabelle 2: Zufallszahlen: Zufallspermutationen der Zahlen von 0 bis 9

Die Zuteilung in Tabelle 1 kann man via Tabelle 2 ganz einfach vornehmen: Man greift sich "blind" einen Zahlenblock heraus. Ignoriert man in diesem Block die Zahlen 0, 7, 8 und 9, da diese in der ersten Teiltabelle in Tabelle 1 nicht benötigt werden, so kann die Zuteilung mit den restlichen 6 Zahlen - die ja eine Zufallspermutation der Zahlen von 1 bis 6 darstellen - vorgenommen werden.

In vielen Computerprogrammen sind Module für Randomisierungen vorgesehen, die neben einer vollständigen Randomisierung auch eine Randomisierung in Blöcken gestatten. Umfassende Hilfe gibt zum Beispiel das Programm **CADEMO** der BioMath GmbH. Mit **BiAS.** sind ebenfalls verschiedene Varianten einer Randomisierung durchführbar.

Kapitel 3:
Deskriptive Statistik

In einer pädiatrischen Studie werden eine Reihe von Laborparametern und anthropometrische Daten (Körperlänge und -gewicht) gemessen. Die Studie umfasst n=70 Neugeborene. Obwohl in den Daten der Urliste zwar sämtliche Information steckt, die in den Ergebnissen nur enthalten sein kann, ist es natürlich zweckmäßig, die Daten zu "verdichten" und - erster Gedanke - vielleicht Mittelwerte o.ä. zu berechnen. Neben einem "Mittelwert" interessiert man sich im Allgemeinen auch für ein Maß für die Unterschiedlichkeit (Variabilität) der Messwerte, und, um die Ergebnisse noch anschaulicher zu machen, für graphische Darstellungen der Daten: Diese Techniken sind Gegenstand dieses Kapitels. Zur Symbolik erinnere man sich, dass in der Statistik der Umfang einer Stichprobe in der Regel mit dem Buchstaben "n" bezeichnet wird. Die eigentlichen Messwerte - n an der Zahl - bezeichnet man mit $x_1,x_2,x_3,...,x_n$. Ein beliebiges x aus dieser Reihe erhält die Bezeichnung x_i, wobei also $1 \le i \le n$ sein muss.

Im Beispiel der pädiatrischen Studie seien die Körpergrößen der n=70 Neugeborenen als chronologische Urliste gegeben:

53	48	45	55	40	51	54	44	48	51
43	55	51	42	48	48	49	50	53	55
55	59	57	48	44	47	48	55	45	44
51	54	46	47	49	51	56	43	51	56
56	50	44	51	56	52	52	49	46	45
41	49	52	51	57	53	47	43	50	52
54	55	51	50	53	52	46	49	50	52

Tabelle 3: Urliste der Körpergrößen von n=70 Neugeborenen

3.1 Maßzahlen der Lage (Mittelwerte)

Mittelwerte oder *Lokalisationsmaße* geben Auskunft über die *Lage der Messwerte* auf der Skala. Tatsächlich gibt es nicht "den Mittelwert", sondern man muss eine ganze Reihe von verschiedenen Mittelwerten unterscheiden:

Bei quantitativen Skalen kommen das arithmetische Mittel, das geometrische Mittel, das harmonische Mittel, der Median und der Modalwert als

"Mittelwert" in Frage. Die letzen beiden Mittelwerte, Median und Modalwert, kann man auch bei Ordinaldaten verwenden, den Modalwert sogar bei Nominaldaten. (Die genannten Skalentypen wurden in Abschnitt 1.1 behandelt.)

Das *arithmetische Mittel* ("Durchschnitt") ist bekanntlich definiert als

$$\bar{x} = \frac{1}{n} \cdot (x_1 + x_2 + ... + x_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Der Durchschnitt wird bei quantitativen Daten berechnet, die sich im Wesentlichen symmetrisch um eben den Durchschnitt gruppieren, also keine "Schiefe" oder auffällig abseits liegende Werte (sogenannte *Ausreißerwerte*) aufweisen. Im Beispiel der Körpergrößen (nicht aber bei z.B. Körpergewichten, vgl. Kapteynsches Gesetz!) kann man davon ausgehen, dass diese Annahmen erfüllt sind und kann als arithmetisches Mittel der Körpergrößen den Durchschnittswert $\bar{x}$ =49.95cm errechnen. Physikalisch kann man das arithmetische Mittel auch als "Schwerpunkt" der Messwerte auffassen.

Das *geometrische Mittel* wird man berechnen, wenn die Messwerte aus relativen Änderungen bestehen: Dies sind zum Beispiel Wachstums- oder Zuwachsraten, Produktionssteigerungen, mittlere Arbeits- oder Wartezeiten, Gehaltserhöhungen etc.. Die Definition lautet

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot .... \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

In einer Bakterienkultur vermehren sich innerhalb von fünf Tagen die Bakterien um 2%, 25%, 3%, 4% und am 5. Tag wieder um 2%. Wie groß ist die mittlere Zuwachsrate in Prozent? (In dem Verlauf äußert sich das in der Biologie häufig beobachtete Phänomen, dass das Wachstum langsam beginnt, sich beschleunigt und dann wieder abnimmt.) Das geometrische Mittel errechnet sich zu 4.13%. Das - fälschlich berechnete - arithmetische Mittel ist mit 7.20% deutlich zu groß, nicht zuletzt wegen des Wertes 25%.

Die per Logarithmierung (Anhang A.3!) äquivalente Definition

$$\bar{x}_G = \exp\left(\frac{1}{n} \cdot \sum_{i=1}^{n} \log(x_i) \right)$$

weist auf eine Anwendungsmöglichkeit in Kapitel 5 hin (parametrische Testverfahren), denn mit Hilfe der Logarithmen kann man sogenannte "schiefe" Verteilungen symmetrisieren und schafft damit die Voraussetzung für die sogenannten "parametrischen Verfahren"; Beispiele finden sich in den Kapiteln 4 und 5. Hier werden die Werte per Logarithmierung symmetrisiert und das daraus berechnete arithmetische Mittel wird mit Hilfe der Exponentialfunktion (als Umkehrfunktion des Logarithmus) auf die ursprüngliche Skala retransformiert, um das geometrische Mittel zu erhalten.

Das *harmonische Mittel* wird benutzt, wenn die zu mittelnden Größen prinzipiell als Quotient dargestellt werden können, wobei entweder der Zähler oder der Nenner konstant bleibt. Die Definition lautet

$$\bar{x}_H = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

Beispiele: Mittlere Geschwindigkeit für Teilstrecken, mittlere Bearbeitungszeit, mittlere Dichte von Gasen oder von Flüssigkeiten, mittlere Überlebenszeit etc.. Das klassische Beispiel für das harmonische Mittel ist die Berechnung der durchschnittlichen Geschwindigkeit:

Ein Radfahrer bewältigt die Fahrtstrecke von A nach B mit einer Durchschnittsgeschwindigkeit von 25 km/h, für den Rückweg mit Gegenwind schafft er nur 20 km/h. Die Durchschnittsgeschwindigkeit für die Gesamtfahrt ist 2/(1/20+1/25)=22.2 km/h. Das - fälschlich berechnete - arithmetische Mittel beträgt 22.5 km/h.

Der *Median* kann für quantitative und auch für ordinale Daten berechnet werden. Im zweiten Fall ist eine Berechnung der oben behandelten Mittelwerte ohnehin sinnlos, bei quantitativen Daten ist die Berechnung angebracht, wenn man keine symmetrische Anordnung der Daten erwarten kann (das klassische Beispiel: Einkommensverteilung). Fast alle Laborwerte verteilen sich nicht-symmetrisch, was sich trivialerweise aus der Tatsache ergibt, dass die Werte in der Regel durch den Wert "0" begrenzt sind (Konzentrationen!), die meisten Werte in einem unteren Bereich liegen, größere Werte jedoch vorkommen und große Werte nicht ausgeschlossen sind. Diese Beschreibung einer *rechtsschiefen Verteilung* findet man zum Beispiel bei den Transaminasen GOT und GPT oder bei der Alkalischen Phosphatase (AP) vor, Ausnahmen von dieser Regel - wie die (symmetrische) Verteilung von Leukozyten in Abschnitt 6.3 - sind eher rar. (Als Beispiel für eine ebenfalls seltene *linksschiefe Verteilung* kann zum Beispiel das Gestationsalter oder die Ergebnisse von Klausuren in Punkten genannt werden, die eine typisch linksschiefe Verteilung aufweisen.)

Zur Bestimmung des Medians ordnet man die gemessenen bzw. beobachteten Daten entsprechend der Größe bzw. gemäß ihrer ordinal gegebenen Ordnung an,

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

wobei die Klammerung ( ) der Indices symbolisiert, dass die Reihenfolge der ursprünglichen Stichprobe durch die eben definierte geordnete Folge ersetzt wurde. Bei einer ungeraden Anzahl von Werten ist der Median die mittlere Zahl in der ranggeordneten Folge, ist die Anzahl der Werte gerade, so errechnet sich der Median aus dem Durchschnitt der beiden mittleren Werte in der ranggeordneten Folge. Als Bezeichnung für den Median verwendet man einen beliebigen Buchstaben, etwa x, und schreibt

darüber an Stelle des Querstrichs wie beim Durchschnitt eine Tilde "~" (gelesen: "x Schlange", geschrieben: $\tilde{x}$).

n gerade: $\tilde{x} = (x_{(n/2)} + x_{(n/2 + 1)}) /2$

n ungerade: $\tilde{x} = x_{((n+1)/2)}$

Möchte man den Median der vier bundesdeutschen Einkommen 1500, 1000, 2000 und 5000 (in Euro) berechnen, so ordnet man diese Zahlen, erhält die geordnete Folge 1000, 1500, 2000, 5000 und bildet den Median als Durchschnitt der beiden mittleren Werte mit 1750 Euro. Fügt man zu den vier Zahlen noch eine fünfte hinzu (z.B. 1250), um eine ungerade Anzahl zu erhalten, so ergibt sich der Median als die mittlere Zahl in der geordneten Folge mit 1500 Euro.

Es ist zu erwarten, dass bei einigermaßen zentral bzw. symmetrisch angeordneten Werten der Median und das arithmetische Mittel einigermaßen übereinstimmen. Daraus könnte man - nicht ganz zu unrecht - schließen, dass im Zweifelsfall stets der Median als Mittelwert bestimmt werden sollte und das arithmetische Mittel im Grunde verzichtbar ist. Vor einer späteren, mehr detaillierteren Diskussion ist vorläufig festzuhalten, dass die Differenz $d=\bar{x}-\tilde{x}$ - oder besser standardisiert $d'=(\bar{x}-\tilde{x})/(x_{max}-x_{min})$ - als Grundlage für eine Beurteilung von Symmetrie verwendet werden kann: $d'\approx 0$ weist auf Symmetrie hin, $|d'|>>0$ eher auf Asymmetrie der vorliegenden Stichprobenwerte. Auch dies wird später wieder aufgegriffen.

Der *Modalwert* oder das *Dichtemittel* $\bar{x}_D$ ist definiert als der häufigste Wert einer vorliegenden Stichprobe bei beliebiger Datenqualität. Diese Variante eines Mittelwertes ist für praktische Zwecke nahezu bedeutungslos, da die so definierte Größe oft nicht eindeutig ist und/oder zu sehr von einzelnen Werten bestimmt ist. Bei Vorliegen von verschiedenen, gleich häufigen Werten ist der Modalwert nicht eindeutig definiert.

In Kapitel 2 (Versuchsplanung) wurde bereits das Modell der zufällig, also randomisiert aus der Population entnommenen, repräsentativen Stichprobe thematisiert. Die Berechung etwa des arithmetischen Mittelwertes $\bar{x}$ erhält dadurch einen tieferen Sinn, denn denkt man nur an das Beispiel der Körpergrößen von Neugeborenen, so gibt der aus der Stichprobe errechnete Wert von im Beispiel $\bar{x}$ =49.95cm doch Anhalt darüber, wie groß Neugeborene "im Allgemeinen" bzw. "im Mittel" sind. Das heißt, man möchte von der Stichprobe auf eine Eigenschaft der Grundgesamtheit schließen, nämlich von dem Mittelwert der Stichprobe auf den Mittelwert der Population; die zuletzt genannte Größe bezeichnet man üblicherweise mit dem kleinen griechischen Buchstaben μ. Dieser Wert μ ist in praxi gänzlich unbekannt, mit Hilfe von $\bar{x}$ kann lediglich eine Schätzung für μ angeben werden. (Alle deskriptiven Größen wie zum Beispiel $\bar{x}$ oder im nächsten Abschnitt s^2 werden als statistische *Schätzungen* oder *Schätzwerte* für die korrespondierenden Größen μ bzw. σ^2 etc. in der Grundgesamtheit bezeichnet.) Die Frage "Wie *genau* sind diese Schätzungen?" kann erst in Kapitel 4 beantwortet werden, an dieser Stelle kann man lediglich sagen, dass die Schätzungen wohl "*richtig*" sind, da man stets mit repräsentativen Stichproben umgeht, mithin also kein *Systematischer Fehler* ("*Verzerrung*", engl. "*bias*") vorliegt, und dass $\bar{x}$ mit μ sicher umso besser übereinstimmt, je größer n ist. Formal bedeutet dies, dass

$$\lim_{n\to\infty} \bar{x} = \mu$$

3.2 Maßzahlen der Variabilität

Als Maßzahlen für die *Variabilität* ("Streubreite") von Messwerten verwendet man grundsätzlich vier verschiedene Größen, nämlich die *Streuung* bzw. die *Standardabweichung*, die *Quartilen* und die *Spannweite*. Während die Streuung bzw. Standardabweichung nur im Zusammenhang mit dem arithmetischen Mittelwert einen Sinn ergibt, verwendet man in anderen Fällen, speziell im Kontext des Medians, die Quartilen und die Spannweite.

Die Streuung s^2 ist definiert als die "mittlere quadratische Abweichung der Einzelwerte x_i von ihrem Durchschnittswert $\overline{x}$". In einer Formel ausgedrückt bedeutet dies

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (\overline{x} - x_i)^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \right)$$

Die Division durch n-1 (und nicht, wie vielleicht erwartet, durch n) kann man sich heuristisch leicht plausibel machen, eine mathematisch präzisere Begründung findet sich später:

Angenommen, man betrachtet eine kleine Stichprobe zu n=2 Werten. Daraus kann man den Durchschnittswert $\overline{x}$ berechnen, denn dieser wird wiederum zur Berechnung der Streuung benötigt. Ist der Durchschnitt gegeben, so stellt man fest, dass in der Formel für s^2 tatsächlich nur eine Größe, x_1 oder x_2, frei variierbar ist: Wählt man die eine, so muss man zwangsläufig die andere vermöge der Beziehung $\overline{x} = (x_1+x_2)/2$ errechnen, denn $\overline{x}$ ist bei dieser Überlegung ja fest gegeben. Da es somit auch für n>2 in der obigen Summe $\Sigma_i(\overline{x} - x_i)^2$ nur n-1 "zufällig variierbare Größen" gibt, wird letztlich auch nur der Mittelwert von n-1 zufälligen Größen gebildet.

Die Anzahl n-1 der frei variierbaren Größen wird als *Freiheitsgrad* bezeichnet und häufig mit fg oder auch mit df ("degree of freedom") abgekürzt. Der Begriff "Freiheitsgrad" spielt in der sogenannten "parametrischen Statistik" eine besondere Rolle, so dass hierauf - speziell in den beiden Kapitel 4 und 5 - immer wieder Bezug genommen wird.

Die *Standardabweichung s* als die "mittlere Abweichung der Einzelwerte vom Durchschnittswert" ist definiert als die Quadratwurzel aus der Streuung s^2. Eine anschauliche und "praktische" Interpretation findet sich in Abschnitt 4.2 ("Gauß-Verteilung"); es wird sich zeigen, dass der Bereich von $\overline{x} - s$ bis $\overline{x} + s$ etwa 2/3 aller Stichprobenwerte einschließt. Diese Angabe gilt nach dem oben gesagten nur für angenähert symmetrisch angeordnete Stichprobenwerte (in Abschnitt 4.2 wird sich herausstellen: streng genommen nur für Gauß-Verteilungen!), in allen anderen Fällen ist die Berechnung der Standardabweichung s nicht sinnvoll.

Die mit s bzw. s^2 korrespondierenden Größen in der Grundgesamtheit heißen Standardabweichung σ und Varianz σ^2. (Etwas präziser: s^2 ist der Schätzwert für σ^2, und für $n \to \infty$ ist der Grenzwert $\lim s^2 = \sigma^2$) Auch hier werden, wie in der Statistik üblich, für die Parameter der Grundgesamtheit

wieder griechische Buchstaben reserviert. Für die folgenden Größen SEM und CV verwendet man im Allgemeinen keine speziellen Bezeichnungen für ihre Populationspendants; diese könnte man naheliegenderweise vielleicht mit $\sigma_{\bar{x}}$ (SEM) und σ/μ (CV) bezeichnen:

Eine problematische Größe ist der *Standardfehler des Durchschnitts* (nach dem englischen Term *Standard Error of the Mean* oft abgekürzt als *SEM* bezeichnet), der definiert ist mit

$$\text{SEM} = s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Damit beschreibt man nicht die Variabilität der ursprünglichen Messwerte x, sondern die Variabilität von Durchschnittswerten $\bar{x}$. Dies erhält einen Sinn, wenn man sich vorstellt, dass man eine Untersuchung sehr oft wiederholen könnte - z.B. N mal - und aus den N Durchschnitten die Standardabweichung $s=s_{\bar{x}}$ dieser Durchschnittswerte berechnen würde. Das Resultat ist offenbar ein Maß für die "Genauigkeit" des Durchschnitts bzw. der Durchschnitte. (Die Konstruktion dieser Größe kann in Kapitel 4 durch die aufschlussreichere Methodik der "Konfidenzintervalle" ersetzt werden.) Wie für die Standardabweichung s der Stichprobenwerte gilt auch hier, dass $s_{\bar{x}}$ nur bei einigermaßen symmetrisch angeordneten, quantitativen Messwerten berechnet werden sollte. Da die Größe $s_{\bar{x}}$ direkt vom Stichprobenumfang n abhängt (und damit von Studie zu Studie, die ja in der Regel unterschiedliche Stichprobenumfänge aufweisen, nicht vergleichbar ist), muss deren Nutzen - obwohl der SEM in der Literatur gerne verwendet wird - als eher zweifelhaft beurteilt werden.

Der *Variationskoeffizient* (engl.: *Coefficient of Variation* CV) ist definiert als der Quotient von Standardabweichung s und arithmetischem Mittel $\bar{x}$:

$$\text{CV} = \frac{s}{\bar{x}}$$

Der Variationskoeffizient ist zweckmäßig, wenn man die Standardabweichung s an der mittleren Größenordnung der Stichprobenwerte relativieren möchte: Häufig wird s bei großen Werten $\bar{x}$ - vielleicht bedingt durch die Messgenauigkeit, aber auch bedingt durch die biologische Variabilität – "groß", so dass man hiermit eine Vergleichbarkeit zwischen im Mittel "kleinen" und "großen" Werten bzw. deren Streuung herstellen kann. Die Größe CV wird gelegentlich auch als *relativer Variationskoeffizient* ausgedrückt ($\text{CV}[\%]=100\%\cdot\text{CV}/\sqrt{n}$) und nimmt Werte zwischen 0% und 100% an (vgl. dazu auch das Buch von Lothar Sachs (2004)).

Die Angaben $\bar{x}$, s, SEM etc. sind sehr weit verbreitet und werden häufig kritiklos auch bei Stichproben errechnet, bei deren Verteilung auch nicht näherungsweise von Symmetrie gesprochen werden kann. In der Medizin

sind solche schiefen Verteilungen der Daten aber eher die Regel als die Ausnahme, so dass man sich alternative Methoden wünscht. Diese sind recht einfach zu definieren und gelten sogar nicht nur für quantitative Daten, sondern auch für Ordinaldaten.

Das denkbar einfachste Maß für die Variabilität heißt *Variationsbreite* oder auch *Spannweite* (im englischen Sprachgebrauch: *Range*) und umfasst den Bereich vom kleinsten Wert $x_{(1)}$ bis zum größten Wert $x_{(n)}$:

$$R = x_{(n)} - x_{(1)} = x_{max} - x_{min}$$

Da diese Größe offensichtlich nur von den beiden Extremwerten $x_{min}=x_{(1)}$ und $x_{max}=x_{(n)}$ abhängt, ist die Angabe im Allgemeinen etwas "unsicher" und sollte deshalb nur in begründbaren Ausnahmefällen Verwendung finden (z.B. wenn n sehr klein ist).

Einen sehr flexiblen Zugang zur Beschreibung von "Variabilität" findet man über sogenannte *Percentilen* (auch: *Quantilen*). Auf Grundlage einer Stichprobe zu n Werten x_i bildet man wie zur Berechnung des Medians die ranggeordnete Folge $x_{(i)}$: i=1,2,...,n (Konvention der Klammerung des Indexes i beachten: Rangordnung!) und ergänzt diese Zahlen durch zwei neue, "künstliche" Werte $x_{(0)}$="-∞" und $x_{(n+1)}$="+∞". Dadurch ergibt sich eine Einteilung der Zahlengerade in n+1 Abschnitte (eine Zahl teilt die Zahlengerade in zwei Teile, zwei Zahlen in drei Teile etc.). Man kann mathematisch zeigen (was allerdings recht aufwendig ist: angesprochen sind Tukey's "statistisch äquivalente Blöcke"), dass diese n+1 "Blöcke" völlig gleichberechtigt sind in dem Sinne, dass jedem Block die gleiche Wahrscheinlichkeit - nämlich 1/(n+1) - zukommt.

Mit diesen Definitionen entspricht beispielsweise die 10%-Percentile (für andere Anteile gilt natürlich Analoges) der Grenze, unterhalb derer 10% und oberhalb derer 90% aller "Blöcke" liegen. Ist die im Beispiel gemäß 10%·(n+1) errechnete Anzahl der Blöcke nicht ganzzahlig, so muss man anteilmäßig linear interpolieren, wie ein einfaches Beispiel zeigt: Der Stichprobenumfang sei n=19, die bereits ranggeordneten Zahlen $x_{(i)}$ seien 35, 40, 41, 44, 52,...,99. Bei n=19 Zahlen ergeben sich offenbar n+1=20 Blöcke, 10% davon sind gerade 2 Blöcke, womit die 10%-Percentile gleich $x_{(2)}$=40 ist. Für n=20 Zahlen würde sich die Anzahl der Blöcke mit 10%·(n+1) = 10%·21 = 2.1 ergeben, so dass die gesuchte Percentile nach Interpolation um den Betrag $0.1\cdot(x_{(3)}-x_{(2)})$ größer wäre als im Fall n=19.

In vielen Lehrbüchern wird eine alternative Percentilenberechnung beschrieben, die auf der "empirischen Verteilungsfunktion" bzw. auf "kumulativen Häufigkeiten" beruht (vgl. dazu auch Abbildung 9 weiter unten): Zu einem gewünschten Anteil p ergibt sich die entsprechende Percentile aus der plausiblen Forderung, dass der Anteil p der Stichprobe *kleiner oder gleich* und gleichzeitig der komplementäre Anteil 1-p *größer oder gleich* dieser Percentile sein muss, wodurch ebenfalls eine symmetrische Definition erreicht wird. Im zweiten Beispiel des vorhergehenden Abschnitts (n=20) liegt nach dieser Definition die 10%-Percentile zwischen den beiden Werten $x_{(2)}$=40 und $x_{(3)}$=41. Die numerischen Unterschiede zwischen den beiden

Varianten sind, speziell für größere n, im Allgemeinen eher geringfügig. Weitere Berechnungsvarianten finden sich in der Literatur, werden aber hier nicht dargestellt.

Das bereits erwähnte Programm **BiAS.** verwendet wie u.a. auch das statistische Standard-Programmpaket **SPSS** die erste Definition mit Hilfe von „statistisch äquivalenten Blöcken“.

Neben den 10%- (und symmetrisch: 90%-) Percentilen findet man häufig auch die 25%- und 75%-Percentilen. Der Bereich dieser beiden Größen $Q_1=x_{(0.25\cdot(n+1))}$ und $Q_3=x_{(0.75\cdot(n+1))}$ umfasst den inneren Bereich von 50% der Stichprobenwerte. Die Bezeichnungsweise Q_1 und Q_3 ist von dem Begriff *Quartile* abgeleitet, was soviel wie "Viertel" - der Stichprobenwerte also – bedeutet; der Bereich von Q_1 bis Q_3 wird auch als *Interquartilbereich* bezeichnet. Zwischen Q_1 und Q_3 befindet sich der Wert Q_2: Q_2 im Sinne von zwei Viertel der Stichprobenwerte kann nach der Definition "mittlere Wert in der ranggeordneten Folge" nur den Median bedeuten.

3.3 Box-Plots

Bei der deskriptiven Auswertung von empirischen Daten sollte man besonderen Wert auf graphische Darstellungen legen. Einige der in den Abschnitten 3.1 und 3.2 definierten deskriptiven Größen kann man - wie in der Literatur üblich - zweckmäßigerweise in einem sogenannten *Box-Plot* zusammenfassen, um auf diese Weise einen ersten Überblick über die Datensituation zu erhalten. Zur Illustration dienen die oben aufgelisteten Daten der Körperlängen von Neugeborenen, für die mit dem Programmpaket **BiAS.** zunächst alle gängigen statistischen Schätzgrößen (gelegentlich auch: "Kenngrößen") errechnet wurden:

Nr.	Name	n	Xquer	s	Median	1.Quartil	3.Quartil
			SEM	CV	Range	Minimum	Maximum
1	Größe	70	49.95	4.34	50.50	47.00	53.00
			0.52	0.09	19.00	40.00	59.00

Abbildung 5: Deskriptive Größen zur Körperlänge von Neugeborenen

Eine graphische Darstellung des geometrischen und des harmonischen Mittels, des Modalwertes, des Variationskoeffizienten CV und des SEM wird in Box-Plots üblicherweise nicht vorgenommen. Eine allgemein verbindliche Darstellungsform gibt es nicht; so symbolisieren manche Programme unterschiedliche Stichprobenumfänge durch unterschiedliche Breiten der "Kästen". Das Programm **BiAS.** erlaubt zusätzlich eine Darstellung der einzelnen Messwerte, die jedoch in Abbildung 6 nicht dargestellt werden.

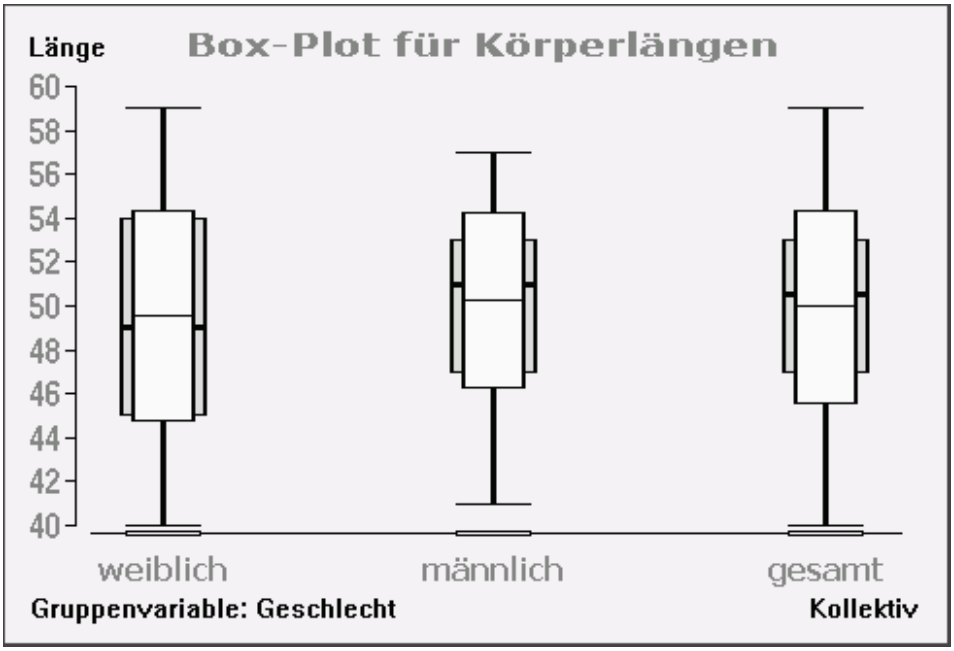


Abbildung 6: Box-Plot für die Körperlängen von Neugeborenen

In Abbildung 6 erkennt man pro Gruppe (weiblich, männlich und gesamt) nicht nur eine Box (engl. Box=Kasten), sondern zwei: Die etwas breitere, hier verdeckte Box umfasst den Bereich von der ersten Quartile Q_1 bis zur dritten Quartile Q_3, den sogenannten Interquartilbereich. Der Median der Stichprobenwerte wird durch eine dicke, horizontale Linie in dieser Box dargestellt. Der etwas dünnere Querstrich entspricht dem Durchschnitt $\bar{x}$, die schmalere Box den Werten $\bar{x}$-s und $\bar{x}$+s. Die beiden "Antennen" markieren die beiden Extremwerte $x_{min}=x_{(1)}$ und $x_{max}=x_{(n)}$. SEM, CV etc. werden üblicherweise nicht graphisch dargestellt.

Box-Plots können eine oder auch mehrere Gruppen bzw. Teilstichproben (stratifizierte Stichproben) darstellen; in Abbildung 6 wurde das Kollektiv der Neugeborenen als Gesamtstichprobe und getrennt nach Geschlecht dargestellt. Eine analoge Darstellung ist auch bei Zeitreihen nützlich; dabei verbindet man die Durchschnitte oder die Mediane durch Linien. Solche Zeitreihendarstellungen finden sich zum Beispiel bei wiederholten Messungen von Patienten im Zeitverlauf.

Box-Plots sind eindeutig sogenannten *Bar-Plots* vorzuziehen, die von vielen Programmpaketen angeboten werden. Bar-Plots stellen in der Regel den Durchschnitt durch eine Säule und die Standardabweichung s nach oben durch einen aufgesetzten Strich dar: Diese Darstellungsform ist einerseits unzureichend, andererseits stiftet sie leicht Verwirrung, da eine Darstellung mit Säulen konventionellerweise für Histogramme (vgl. dazu Abschnitt 3.5) reserviert ist. Von einer Verwendung von Bar-Plots ist also eher abzuraten, womit sich nachdrücklich eine Darstellung von Box-Plots wie in Abbildung 6 empfiehlt.

3.4 Relative Häufigkeiten

Relative Häufigkeiten finden sich überall im täglichen Leben ("wieviel Prozent" hatte die Partei XY bei der letzten Bundestagswahl?) und müssten in diesem Skript vielleicht gar nicht eigens erwähnt werden, wenn nicht die Statistik relative Häufigkeiten in eigener Weise behandeln würde.

Die *relative Häufigkeit* h (man erinnere sich an Abschnitt 0.1!) ist definiert als der relative Anteil einer Merkmalsausprägung am Stichprobenumfang n. Die relative Häufigkeit nimmt wie die Wahrscheinlichkeit ausschließlich Werte zwischen 0 und 1 an. Die nicht am Stichprobenumfang relativierte, also absolute Anzahl wird *absolute Häufigkeit* genannt.

Im Würfelspiel kann man zählen, wieviele Sechsen bei n=2000 Würfen vorgekommen sind, vielleicht n_6=312. Die relative Häufigkeit für die Ausprägung "6" errechnet sich also mit 312/2000=0.156. Erwartet hätte man 1/6=0.167, da die "6" ja eine von sechs verschiedenen Zahlen auf dem Würfel ist, die im Wesentlichen gleich oft vorkommen sollten. Damit vergleicht man also die beobachtete relative Häufigkeit h=0.156 mit der Wahrscheinlichkeit P=1/6, beim Würfeln eine Sechs zu würfeln: Ist die Abweichung noch "zufällig" oder nicht? Ist der Würfel womöglich "gezinkt" und bringt weniger Sechsen als theoretisch zu erwarten sind? (Vgl. auch Abschnitt 0.1, Vergleich der "logischen" und der "frequentistischen" Wahrscheinlichkeitsdefinition.)

Auf eine Entscheidungsstrategie zu dieser Frage kann erst im Rahmen der Hypothesenprüfung in Abschnitt 5.1 eingegangen werden. Hier kann man nur feststellen, dass relative Häufigkeiten im Sinne des letzten Abschnittes als Schätzwerte für die korrespondierenden Wahrscheinlichkeiten P aufzufassen sind. Stellt man sich vor, dass der Stichprobenumfang n immer weiter vergrößert wird, also immer weiter gewürfelt wird, so wird h "irgendwann" mit P übereinstimmen, formal:

$$\lim_{n\to\infty} h = P$$

Ad hoc fallen dazu viele Beispiele aus der Medizin ein, etwa der Anteil "Rhesus-Faktor positiv Rh+" in der Gesamtbevölkerung oder andere, Medizin-typische Beispiele wie spezielle relative Häufigkeiten, die auch als *Krankheitsstatistiken* bezeichnet werden. Zur Ermittlung dieser Krankheitsstatistiken wird eine bestimmte Anzahl N von Personen über z.B. ein Jahr beobachtet (z.B. die Gesamtbevölkerung der BRD, N ist die mittlere Anzahl Personen) und es werden in diesem Zeitraum vier weitere charakteristische Häufigkeiten ermittelt: n_a ist die Anzahl der Personen, die zu Beginn des Beobachtungszeitraums erkrankt waren. n_k ist die Anzahl im Zeitraum neu erkrankter Personen. n_e ist die Anzahl der am Ende des Zeitraumes gerade erkrankten Personen. n_g ist die Anzahl der im Zeitraum gestorbenen Personen.

Die erforderlichen Definitionen der Krankheitsstatistiken sind nun recht einfach und beziehen sich alle auf den definierten Beobachtungszeitraum (in der Regel ein Jahr) und den mittleren Kollektivumfang N in diesem Zeitraum (hier nach Ramm und Hofmann (1975)):

Inzidenz	=	n_k / N	Neuerkrankungsrate im Beobachtungszeitraum
Prävalenz	=	n_a / N	Erkrankungsrate zu Beginn des Beobachtungszeitraums
Mortalität	=	n_g / N	Sterberate im Beobachtungszeitraum
Morbidität	=	n_k / N	Erkrankungsrate (i.A. identisch mit Inzidenz)
Letalität	=	$n_g / (n_a+n_k-n_e)$	Sterberate unter den Erkrankten

Es ist zu beachten, dass - von Epidemien abgesehen - i.d.R. $n_a=n_e$ ist und die oben dargestellte mathematische Definition der Letalität auch vereinfacht dargestellt werden kann; nach den Rechenregeln für Wahrscheinlichkeiten (vgl. Kapitel 0.3) ist dann auch entsprechend vereinfacht Mortalität = Morbidität $\times$ Letalität. In Abschnitt 6.4 ("Bewertung diagnostischer Tests") finden sich einige weitere spezielle relative Häufigkeiten (Spezifität, Sensitivität, Prädiktive Werte etc.), die im Labor und in der Diagnostik häufig verwendet werden.

3.5 Histogramme

Histogramme kann man grundsätzlich für alle Skalentypen darstellen. Als einfaches Beispiel bietet sich die bereits verwendete Körpergröße von Neugeborenen an, die in Gestalt ihrer relativen Häufigkeiten deskriptiv ausgewertet und in Abbildung 7 als diskretes Histogramm (griech. Histos = Stab) dargestellt werden kann.

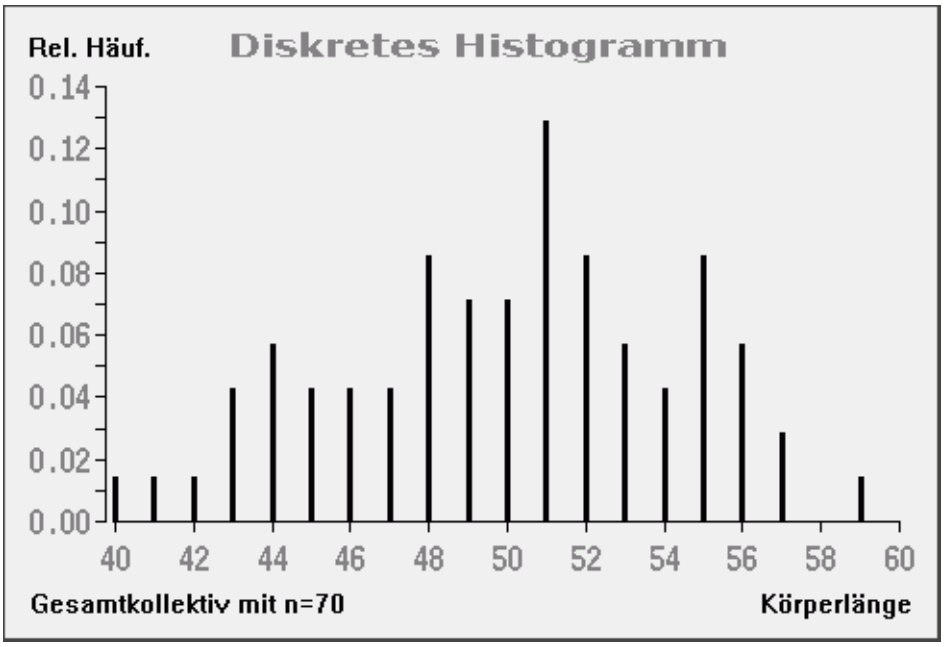


Abbildung 7: Diskretes Histogramm zur Körperlänge Neugeborener

Das Merkmal "Körperlänge" ist an sich stetig, im Beispiel jedoch diskretisiert angegeben. ("Diskretisiert" bedeutet, dass eine an sich stetige Größe bedingt durch die Messgenauigkeit, vielleicht aber auch dadurch, dass eine Angabe von weiteren Dezimalstellen uninteressant oder sinnlos ist, nur auf eine beschränkte Anzahl Stellen genau angegeben wird - im Beispiel der Körpergrößen auf einen Zentimeter genau.) Der diskretisierten Skala entsprechend wird deshalb in Abbildung 7 ein diskretes Histogramm dargestellt. Die Länge eines Stabes bzw. Striches ist proportional zu der relativen Häufigkeit der entsprechenden Stufe der diskreten Skala. Die Bereiche zwischen den Skalenstufen bleiben natürlich frei, da dort keine Werte beobachtet werden können.

Interpretiert man die Körperlänge - der Sachlage eher angemessen - als stetige Variable, so erhält man eine etwas andere Darstellung, die in Abbildung 8 wiedergegeben ist.

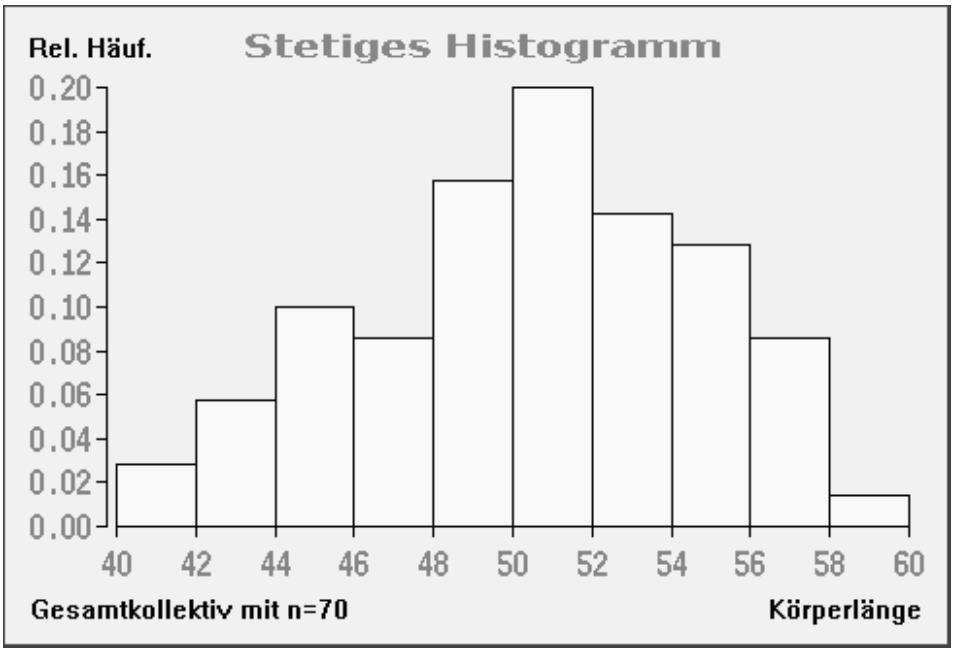


Abbildung 8: Stetiges Histogramm zur Körperlänge Neugeborener

Die Darstellung in Abbildung 8 erfolgt nicht ganz willkürlich, aber auch nicht ganz willkürfrei. Als Faustregel sollte man - stetige, quantitative Skala vorausgesetzt - für eine Stichprobe vom Umfang n etwa Wurzel aus n viele Intervalle vorsehen; im Beispiel ist n=70, die Wurzel daraus beträgt 8.4. Ein Intervall sollte also etwa $(x_{(n)}-x_{(1)})/\sqrt{n}$ =(59-40)/8.4=2.26 cm lang sein, zweckmäßig gerundet sind das 2 cm. Daraus ergeben sich 10 Intervalle. Man zählt ab, wieviele Werte in die verschiedenen Intervalle fallen, errechnet aus diesen absoluten Häufigkeiten die relativen Häufigkeiten und trägt über jedem Intervall einen "Balken" proportional zur relativen Häufigkeit ab. Natürlich muss man definieren, ob z.B. 44 in das

Intervall von 42 bis 44 oder in das benachbarte von 44 bis 46 fallen soll: Im vorliegenden Histogramm würde der Wert 44 in das obere Intervall fallen. Diese Darstellungsform nennt man gelegentlich auch *Balkendiagramm*. (Es ist zu beachten, dass bei diesem – stetigen! - Histogramm auch die Bereiche zwischen den Skalenpunkten beteiligt sind, da die Skala jetzt als stetig ("kontinuierlich") aufgefasst wird. Die Intervalldefinition wird durch die Schreibweise – Vorsicht: runde und eckige Klammern! - z.B. $44 \in (42,44]$ bzw. $42 \notin (42,44]$ symbolisiert.)

Manchmal werden Balkendiagramme in einer Graphik neben- oder übereinander dargestellt, um verschiedene Gruppen wie Geschlecht, Altersgruppen etc. miteinander zu vergleichen. Die Methode ist prinzipiell die gleiche wie oben beschrieben. Quasi-dreidimensionale Histogramme mit perspektivisch dargestellten Säulen, die von vielen Programmpaketen erzeugt werden, sind zwar graphisch ansprechend, aber sachlich ungünstig, da sie in aller Regel die Größenverhältnisse nur verzerrt wiedergeben und deshalb besser zu meiden sind.

Balkendiagramme kann man auch für nicht-quantitative Daten verwenden. Das Prinzip ist natürlich wieder das gleiche wie oben und muss deshalb nicht mehr weiter ausgeführt werden.

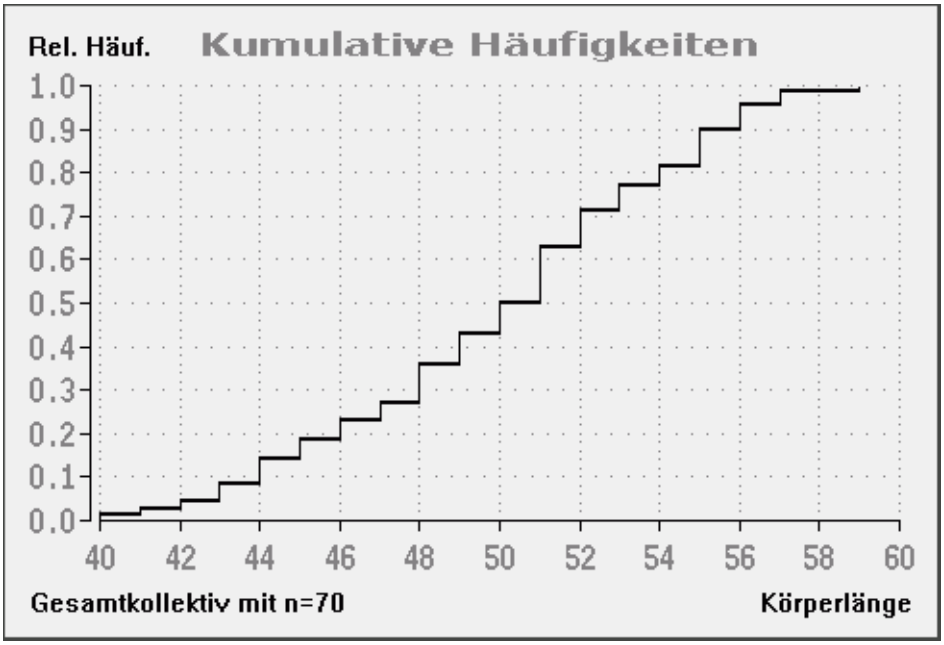


Abbildung 9: Kumulative Häufigkeiten zur Körperlänge Neugeborener

In manchen Situationen können quantitative Daten besser anhand eines *Kumulativen Histogrammms* (auch: *Summenkurve* oder *empirische Verteilungsfunktion*) beurteilt werden. Bezeichnet man die relativen Häufig-

keiten von N Intervallen mit $h_1,h_2,...,h_N$, so werden an Stelle der relativen Häufigkeiten h_i die Werte der *Relativen Häufigkeitssumme* H_i aufgetragen:

$$H_i = \sum_{j=1}^{i} h_j = h_1 + h_2 + \ldots + h_i$$

Noch günstiger (weil feiner aufgelöst) kann man die aufzutragende Kurve bei 0 beginnen und erhöht den Wert für i=1,2,...,n jeweils an der Stelle $x_{(i)}$ um den Betrag 1/N. Dies ergibt Abbildung 9.

Die unterschiedlichen Höhen der Stufen sind dadurch begründet, dass einzelne Messwerte zwei- oder mehrfach vorkommen. Diese *Bindungen* (engl. *ties*) führen dazu, dass die Kurve nicht nur um 1/N, sondern, je nach Anzahl gleicher Werte wie in Abbildung 9, gelegentlich um 2/N, 3/N oder um einen anderen Betrag erhöht wird.

3.6 Kreisdiagramme

Kreisdiagramme – bevorzugt für nominale bzw. kategoriale Daten geeignet – sind jedem aus der Tagespresse geläufig, so dass sicher wieder ein einfaches Beispiel genügt, um das Prinzip darzustellen. Die nächste Abbildung 10 zeigt die prozentuale Verteilung der Analyseaufträge für den Autoanalyser in einem Labor:

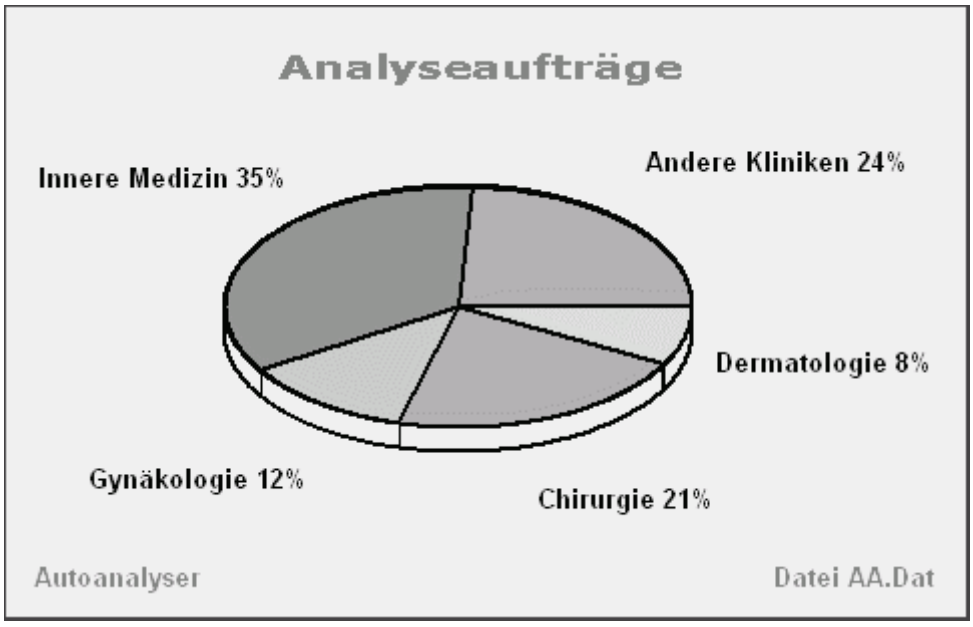


Abbildung 10: Kreisdiagramm für die Analyseaufträge eines Labors

In einem Labor werden die Analyseaufträge für den Autoanalyser dokumentiert. Aus der Inneren Medizin kamen 35% der Aufträge, aus der Gynäkologie 12%, aus der Chirurgie 21% und aus der Dermatologie 8%. Die übrigen 24% der Aufträge kamen aus anderen Kliniken; die Prozentzahlen addieren sich zu 100%. Überträgt man dies auf den Kreis, so werden die vorliegenden Prozentzahlen P_i auf Kreissektoren φ abgebildet durch $\varphi_i = (P_i \cdot 360/100)°$. Die Summe dieser Winkel ergibt natürlich 360°.

Kreisdiagramme können prinzipiell auch als Balkendiagramme dargestellt werden, wobei die Methode eher den persönlichen Gepflogenheiten überlassen bleibt. Ein Nachteil der Kreisdiagramme ist ganz sicher, dass keine vergleichenden Darstellungen wie bei Histogrammen oder bei Box-Plots möglich sind. Bei quantitativen Daten sind Kreisdiagramme nicht sinnvoll, so dass eine Anwendungsmöglichkeit ausschließlich bei qualitativen, speziell bei kategorialen Daten gegeben ist. Perspektivisch dargestellte Kreisdiagramme führen durch die dreidimensionale Darstellung oft zu einer Überbetonung des Bildvordergrundes und damit möglicherweise zu einer optischen Verzerrung, so dass damit Vorsicht geboten ist.

3.7 Scattergram

Unter einem *Scattergram* oder *Streudiagramm* (auch: *XY-Plot*, synonym, aber weniger gebräuchlich: *Punktwolke*) versteht man eine zweidimensionale Darstellung von zwei Variablen:

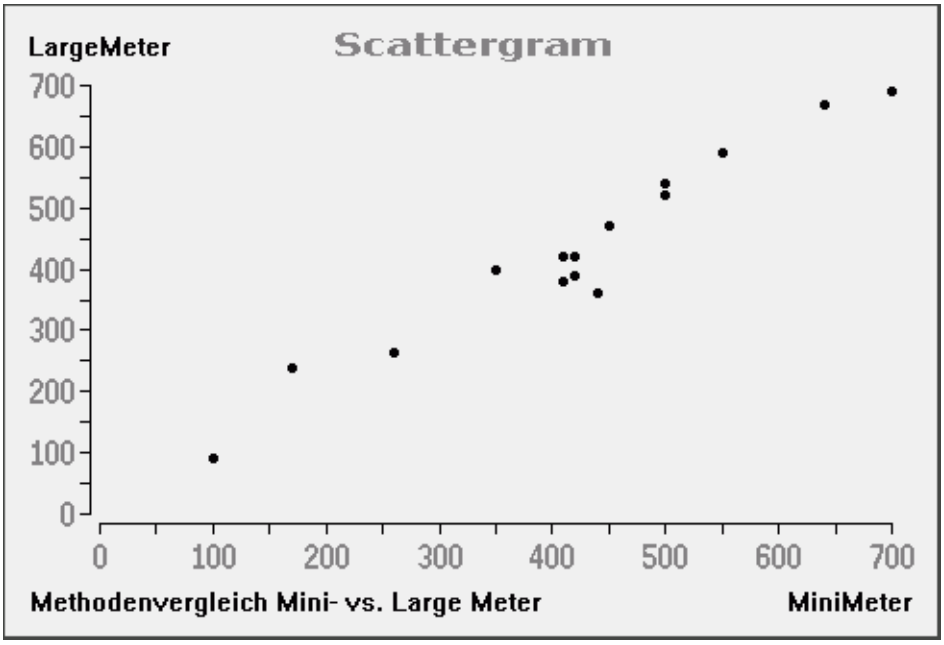


Abbildung 11: Scattergram zum Vergleich Mini- versus Large-Meter

In einem Scattergram wird die eine Variable längs der Abszisse ("X-Achse") und die andere längs der Ordinate ("Y-Achse") abgetragen. Solche Darstellungen findet man zum Beispiel im Labor beim Vergleich von zwei Messmethoden bzw. Messgeräten oder in der Schilddrüsendiagnostik bei der Darstellung zweier Hormonwerte (etwa T3 und T4, vgl. Kapitel 5, Abbildung 22!). Die Darstellung wird in gewissen Fällen um eine sogenannte "Regressionsgerade" erweitert – Einzelheiten dazu finden sich in Kapitel 5.

Im Beispiel wurde eine Laborbestimmung am gleichen Material mit jeweils zwei Geräten vorgenommen (Mini-Meter und Large-Meter). Die Zahlenwerte sind sicher weniger aufschlussreich als die graphische Darstellung der Ergebnisse in Abbildung 11.

Im speziellen Beispiel von Abbildung 11 würde man sicher gerne noch die Winkelhalbierende einzeichnen, um sich einen besseren Eindruck von der Güte der Übereinstimmung der beiden Geräte zu verschaffen. Diese erste graphische Beurteilung wird weiter unten durch eine exakte, objektive Auswertungmethodik präzisiert, so dass auch hier wieder auf einen späteren Abschnitt (diesmal auf Abschnitt 6.2, "Vergleich von Analysemethoden") hingewiesen wird.

3.8 Zeitverläufe

Ein häufiges graphisches Problem ist die Darstellung von Daten im Zeitverlauf. Abbildung 12 wurde mit **BiAS.** erzeugt und zeigt die Darstellung eines Lungenfunktionsparameters im Vergleich prä-, intra- und post-OP. In der Abbildung sind die *individuellen* Zeitverläufe dargestellt:

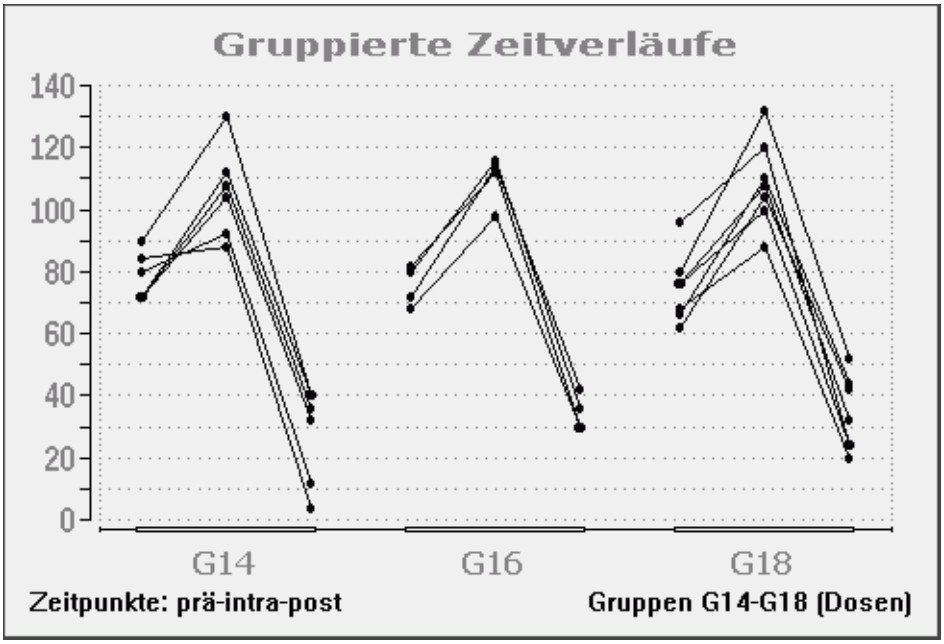


Abbildung 12: Zeitverläufe individueller Patienten

Die Daten in Abbildung 12 sind bezüglich dreier verabreichter Dosen (G14-G18) gruppiert, natürlich ist eine vergleichbare Darstellung auch für nur eine Behandlungsgruppe möglich.

Das Programm **BiAS.** sieht zahlreiche Varianten von Zeitreihendarstellungen vor, zum Beispiel – wie Abbildung 13 zeigt – Zeitverläufe mit Box-Plots (mit Median und Quartilen oder wahlweise mit Durchschnitt und Standardabweichung) oder analog mit Bar-Plots. In allen Graphiken können die individuellen Werte eingeblendet werden.

Die nächste Abbildung 13 wurde mit den gleichen Daten wie Abbildung 12 erzeugt und zeigt – ebenfalls gruppiert - die Zeitverläufe mit Box-Plots:

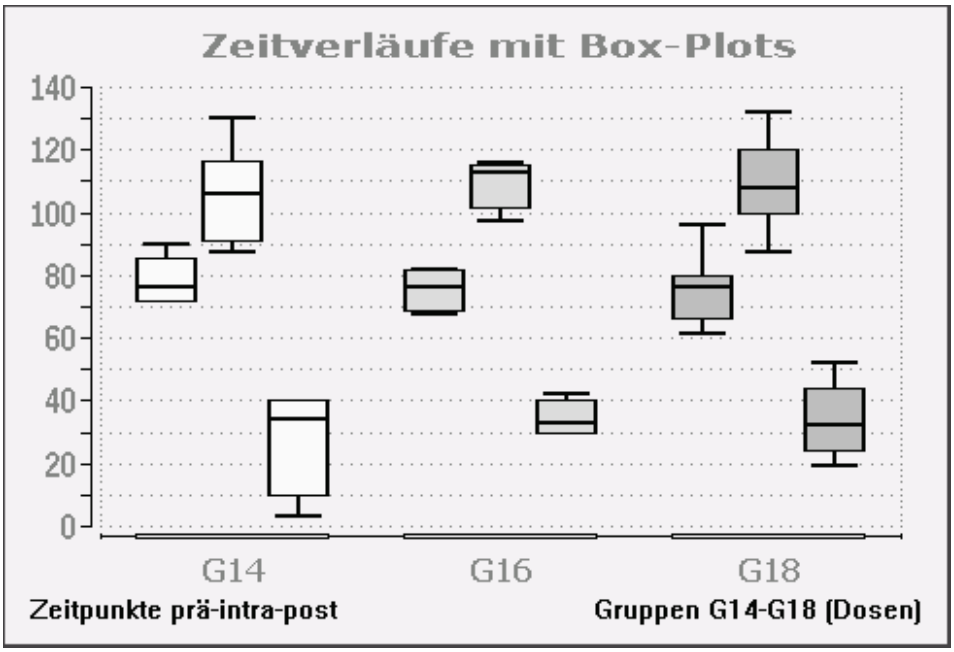


Abbildung 13: Zeitverläufe mit gruppierten Box-Plots

3.9 Ausblick

Neben den eben beschriebenen Verfahren gibt es einen schier unerschöpflichen Fundus weiterer graphischer Methoden, teilweise auch für dreidimensionale Darstellungen (wenn auch der Nutzen solcher Methoden oft zweifelhaft ist). Mit Hilfe von einschlägigen Computerprogrammen (zum Beispiel auch mit dem Tabellenkalkulationsprogramm **Excel**) sind natürlich noch viele weitere Graphiken denkbar, die Frage ist jedoch in allen Fällen, ob eine Graphik klar, möglichst verzerrungsfrei und eindeutig interpretierbar ist.

Graphik-Programme wie zum Beispiel **Harvard Graphics** (dazu auch Kapitel 7) bieten viele Möglichkeiten zur Erzeugung von brauchbaren Graphiken und Präsentationen, nicht zu vergessen das bekannte Microsoft-Programm **PowerPoint**. Aus genanntem Grund beschränkt sich die Darstellung dieses Skripts auf die Arbeitsgraphiken des Programms **BiAS.**.

Quasi-dreidimensionale Darstellungen von an sich zweidimensionalen Datenstrukturen sollte man nach Möglichkeit vermeiden. Beispiele für einfache zweidimensionale Darstellungen finden sich unter anderem in den Abschnitten 3.5 (Histogramme) und 3.6 (Kreisdiagramme). Flächige Darstellungen sind in aller Regel vorzuziehen, da diese eine verzerrungsfreie Wiedergabe der Größenverhältnisse der dargestellten Datenelemente gewährleisten.

Dreidimensionale Häufigkeitsverteilungen werden üblicherweise in Tabellenform wiedergegeben (zum Beispiel eine zweidimensionale Häufigkeitsverteilung für die beiden Transaminasen GOT und GPT); auf ein konkretes Beispiel wird verzichtet.

Eine dreidimensionale, perspektivische Darstellung von Datenzusammenhängen ist sicher reizvoll und produziert mitunter interessante Graphiken, ist aber gerade wegen der perspektivischen Möglichkeiten unter Umständen auch suggestiv oder auch irreführend. Die nicht willkürfreie Wahl eines Blickwinkels kann einerseits gewisse Eigenarten des Datenmaterials hervortreten, andererseits aber auch verschwinden lassen: Eine andere Wahl der Perspektive hat unter Umständen eine gänzlich andere Interpretation einer Graphik zur Folge.

In den nachfolgenden Kapiteln finden sich vereinzelt noch weitere graphische Darstellungsmöglichkeiten, zum Beispiel in Abschnitt 4.3 die Darstellung eines vergleichenden Histogramms bei zwei Gruppen.

Weitere graphische Darstellungsformen finden sich in nahezu jedem Buch über angewandte Statistik; Hinweise finden sich reichlich in dem Buch von Lothar Sachs, Springer-Verlag 2004.

Das Buch von Linder und Berchtold, Statistische Auswertung von Prozentzahlen (UTB Birkhäuser) beschreibt zahlreiche Möglichkeiten zu der oft problematischen Darstellung von Prozentzahlen.

Als Klassiker der sogenannten *Explorativen Datenanalyse* kann das Buch des amerikanischen Statistikers John W. Tukey (1915-2000), Exploratory Data Analysis, Addison-Wesley 1977 genannt werden. Das Buch (Zitat: "The greatest value of a picture is when it forces us to notice what we never expected to see") enthält zahlreiche, auch zum Teil ungewöhnliche Möglichkeiten der graphischen Darstellung von Daten wie *stem-and-leaf-plots*, *letter-value-display*, einige interessante Techniken bei Datentransformationen und viele andere, natürlich auch konventionelle Darstellungen wie Histogramme und Box-Plots.

Kapitel 4:
Konfidenzintervalle

In diesem Kapitel wird allmählich das Feld der deskriptiven Statistik verlassen und über die Brücke der *Konfidenzintervalle* das neue Gebiet der *induktiven* oder auch *schließenden Statistik* betreten: Alle bisher behandelten deskriptiven Verfahren lassen zwar bereits Schlüsse auf die Grundgesamtheit zu, deren Eigenschaften (Mittelwert, Variabilität etc.) man letztlich in fast jeder Untersuchung kennenlernen will, trotzdem weiß man nicht, wie weit die verwendeten Schätzgrößen möglicherweise von den "wahren" Parametern der Grundgesamtheit entfernt sein können. Weiter oben konnte nach der Definition des arithmetischen Mittelwertes $\bar{x}$ nur festgestellt werden, dass die versuchsplanerischen Vorbereitungen (repräsentative Stichproben etc.) dazu führen müssten, dass $\bar{x}$ als Mittelwert einer repräsentativen Stichprobe wohl im Wesentlichen da liegen sollte, wo sich auch μ als Mittelwert der Grundgesamtheit befindet. In diesem Sinne darf die Schätzung $\bar{x}$ zwar als *richtig* bezeichnet werden, da via Repräsentativität kein systematischer Fehler im Spiel sein kann, aber: Wie *genau* ist denn die Schätzung $\bar{x}$? Wie weit kann $\bar{x}$ denn noch von dem unbekannten "Erwartungswert" μ entfernt liegen? Diese Frage steht im Mittelpunkt dieses Kapitels und wird über den Begriff der statistischen Nullhypothese zu einem ersten Ansatz für das bereits mehrfach angekündigte *statistische Testen* führen.

Bevor diese Zusammenhänge etwas weiter geklärt werden, sollte man sich einen eigenen Standpunkt für die weitere Lektüre verschaffen. Sicher muss man nicht alle mathematischen Einzelheiten dieses Skriptums im Detail erarbeiten, um Statistik anzuwenden oder um Statistik zu verstehen, und so genügt es vielleicht manchem, einen wenigstens intuitiven, anschaulichen Weg zu der Materie zu finden: Hier und im Folgenden wird deshalb stets parallel ein mehr mathematischer und ein mehr intuitiver Zugang verfolgt. Wem weniger an der Mathematik gelegen ist, kann die entsprechenden Passagen auch eher zurückhaltend konsumieren, während anderen Leserinnen und Lesern vielleicht eine mehr formale, ausführlichere Darstellung zum Verständnis nützlich ist. Unabhängig vom eingeschlagenen Weg sollte man die Grundzüge der Methoden verstehen und in der Lage sein, diese richtig zu anzuwenden, richtig zu interpretieren und in konkreten Studiensituationen kritisch einzubringen. Dazu werden noch zahlreiche Beispiele aus der Praxis beitragen.

4.1 Wie genau sind statistische Schätzwerte?

Warum muss man zwischen richtig und genau unterscheiden? Zur Illustration stelle man sich vor, dass man mit Pfeilen auf eine Zielscheibe wirft und "irgendwo" trifft. In Abbildung 14 werden die vier grundsätzlich denkbaren Situationen graphisch dargestellt:

Ein guter Spieler wird richtig und genau werfen. "Richtig" bedeutet, dass er irgendwie um den Scheibenmittelpunkt herum wirft und sozusagen "im Durchschnitt" die Mitte der Scheibe trifft. Dies ist aber auch bei dem zweiten Spieler rechts oben der Fall, nur wirft dieser - im Gegensatz zum ersten - "ungenau", weist also eine viel größere Streuung seiner Treffer auf. Damit erklären sich auch die beiden Situationen in der zweiten Zeile. Der dritte Spieler links unten wirft zwar "genau", weist aber einen erheblichen systematischen Fehler ("*bias*") auf, denn er trifft zwar immer ziemlich "genau" an die gleiche Stelle, aber nur eben an die falsche. Der eher hoffnungslose Fall rechts unten muss sicher nicht weiter erläutert werden.

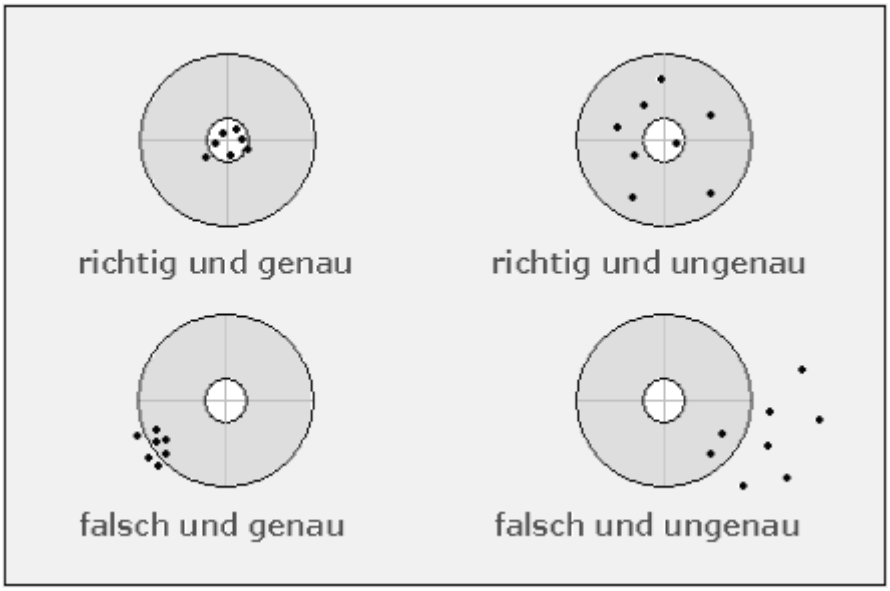


Abbildung 14: Richtig und genau

Überträgt man das Beispiel auf den statistischen Kontext, so ist der Scheibenmittelpunkt mit dem Mittelwert μ der Grundgesamtheit zu identifizieren, während ein Pfeiltreffer einem errechneten Durchschnittswert $\bar{x}$ entspricht. Der "systematische Fehler" des Spielers kann mit dem "systematischen Fehler" verglichen werden, der durch nicht-zufällige Stichproben induziert sein kann. Die "Genauigkeit" von Messungen wird in den nächsten Abschnitten beschrieben, wobei nicht nur die Streuung der Einzelwerte, sondern insbesondere die Streuung der Durchschnittswerte $\bar{x}$ betrachtet wird. Die Frage lautet somit: "Wie genau schätzt (als Pfeil: "trifft") der Durchschnitt $\bar{x}$ als Mittelwert der Stichprobe den unbekannten Parameter μ, also den Mittelwert der Grundgesamtheit?"

Das Beispiel aus Abbildung 14 kann ohne Weiteres auf das praktische Arbeiten zum Beispiel im Labor, in der Radiologie und natürlich auch auf beliebige andere Bereiche übertragen werden. Dabei wird man die "Genauigkeit" z.B. von Messungen am Röntgenbild (z.B. Planimetrie von Tumoren o.ä.) durch Mehrfachmessungen abschätzen, die Frage, ob man auch "richtig" misst, kann man nur durch Messungen an einer quantitativ bekannten

Röntgenaufnahme oder im Vergleich zu einem anderen Messverfahren bzw. zu einem Standard beantworten. Für Laboruntersuchungen gibt es von Firmen vorgegebene Testkits, die man zur Qualitätssicherung und zur Eichung eines Gerätes verwenden kann. Vergleichen Sie dazu bitte auch das Kapitel 6.1 ("Qualitätssicherung") weiter unten.

Bei der Durchführung klinischer Studien ist man darauf angewiesen, auf Grund der Messwerte an einem Patientenkollektiv Aufschluss über die Genauigkeit aller Schätzgrößen zu erhalten, eine häufige Wiederholung einer Studie ist dabei selbstredend ausgeschlossen. Diese Überlegungen werden hier nur modellhaft vorgenommen; eine statistische Analyse ist natürlich ebenfalls ohne Studienwiederholung durchführbar.

Zur Beantwortung aller angesprochenen Fragen fehlt noch ein wichtiger mathematischer Baustein, der im nächsten Abschnitt 4.2 besprochen wird.

4.2 Die Gauß-Verteilung

Die *Gauß-Verteilung* (oder synonym: *Normalverteilung*) ist eine stetige mathematische Funktion (vgl. Anhang A.4), die in Abbildung 15 graphisch dargestellt ist. Diese Funktion, deren Definition aus der Beschriftung der Ordinate ersichtlich ist, hängt von x ab (Abszisse) und besitzt die beiden Parameter μ und σ. Eine Verbindung zu den bereits weiter oben verwendeten Symbolen μ und σ kann, muss aber noch nicht hergestellt werden:

Die von Carl Friedrich Gauß (1777-1855) vollendete und später nach ihm benannte Wahrscheinlichkeitsverteilung geht ursprünglich auf Abraham de Moivre (1667-1754) zurück, der 1711 den Übergang von der Binomial- zur Normalverteilung entdeckte.

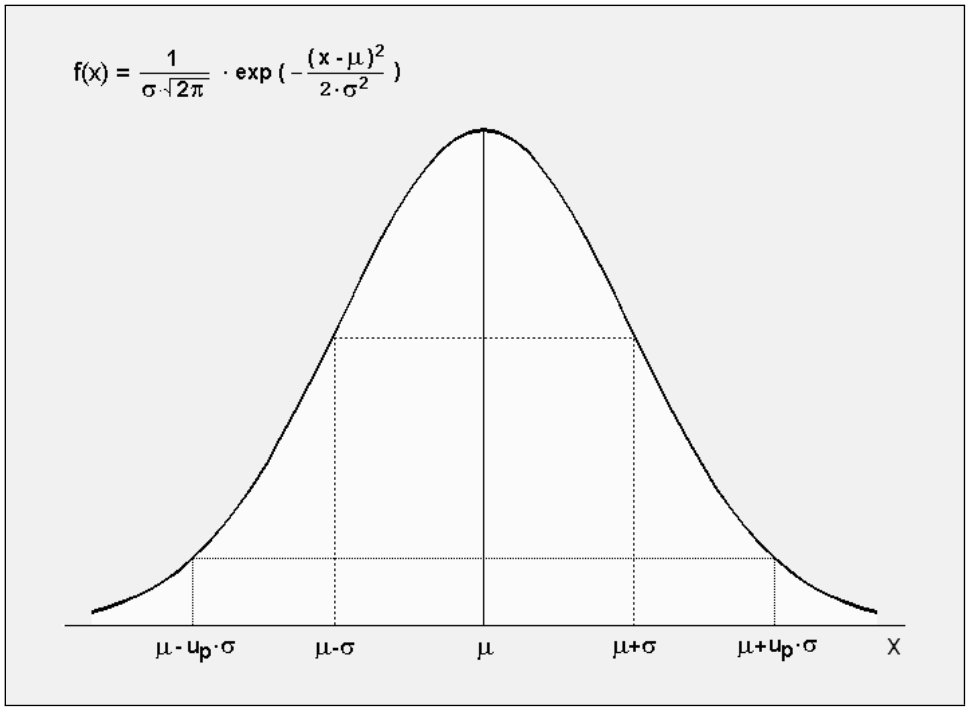


Abbildung 15: Die Gauß-Verteilung ("Normalverteilung")

Wie im Anhang A.5 (Differentialrechnung) nachzulesen ist, kann man per Kurvendiskussion (via erste, zweite und dritte Ableitung der Funktion f(x)) das Maximum und die Wendepunkte der Kurve feststellen. Das Maximum - auch ohne Ableitung am Exponenten der Funktion erkennbar - befindet sich an der Stelle μ, die Wendepunkte, an denen die Kurve, von "links" kommend, von einer Links- in eine Rechtskrümmung und später wieder in eine Linkskrümmung wechselt, befinden sich bei $\mu-\sigma$ und $\mu+\sigma$.

Gemäß Anhang A.6 (Integralrechnung) kann man auch die zwischen der Kurve und der Abszisse eingeschlossene Fläche untersuchen, also das Integral der Funktion bestimmen. Da die Funktion für alle x von $-\infty$ bis $+\infty$ definiert ist, bildet man das bestimmte Integral in den Grenzen von $-\infty$ bis $+\infty$ und stellt fest, dass sich gerade der Wert "1" ergibt: Die Gesamtfläche unter der Kurve beträgt 1, die Funktion ist somit auf 1 normiert (vgl. Definition der Funktion in Abbildung 15, Faktor $1/(\sigma\sqrt{2\pi})$). Für das Integral der Gauß-Verteilung existieren umfangreiche Tabellen wie etwa Tabelle 4.

Andere Werte für das bestimmte Integral sind natürlich ebenfalls denkbar: Interessiert man sich beispielsweise für das Integral bzw. für die Fläche P zwischen den Abszissenwerten $\mu-\sigma$ und $\mu+\sigma$, so erhält man per Integration den Wert P=0.683 oder, prozentual ausgedrückt, einen Anteil von 68.3% der Gesamtfläche unter der Kurve.

Für beliebige Abszissenpunkte $\mu-u\cdot\sigma$ und $\mu+u\cdot\sigma$ kann man natürlich ebenfalls die eingeschlossene Fläche P berechnen. u ist dabei ein beliebiger Faktor (im Beispiel $\mu\pm\sigma$ war u=1, im ersten Beispiel ist - man entschuldige die Schreibweise - u="∞"): Jedem Faktor u - oder jetzt besser: u_P - entspricht offenbar eine gewisse Fläche P, so wie dem Faktor $u=u_{0.683}=1$ die Fläche P=0.683 entspricht. Die folgende Tabelle gibt über weitere Werte Aufschluss und erspart das Integrieren, das ohnehin nicht explizit, sondern nur mit Methoden der Numerischen Mathematik möglich ist.

Gauß-Verteilungen ("*Normalverteilungen*") werden mit $N(\mu,\sigma^2)$ bezeichnet; standardisiert mit $\mu=0$ und $\sigma^2=1$ also mit N(0,1). Tabelle 4 enthält dazu gängige Werte, die auch später im Rahmen des Nullhypothesentestens wieder relevant sind. Üblicherweise - so auch hier - werden in solchen Tabellen nicht die P-Werte im Sinne der Flächen, sondern deren komplementäre Werte $\alpha=1-P$ angegeben. Zur Orientierung wurde in der 3. Zeile der Tabelle 4 der Faktor $u_{0.683}$ eingefügt, dem bei der Gauß-Verteilung eine Fläche von 0.683=68.3% entspricht: Die zugehörige komplementäre Größe $\alpha=1-0.683=0.317$ ist in der Spalte "zweiseitig" aufgelistet. "Zweiseitig" bedeutet, dass man, wie oben beschrieben, um μ symmetrische Grenzen berechnet. "Einseitig" ist so zu verstehen, dass die obere Integrationsgrenze zwar bei $\mu+u_P\cdot\sigma$, die untere aber nicht bei $\mu-u_P\cdot\sigma$, sondern bei "$-\infty$" liegt. Ein Vergleich der Spalten "einseitig" und "zweiseitig" macht die Bedeutung der Schreibweisen offensichtlich.

u_P	$\alpha = 1 - P$ zweiseitig	einseitig
0.674490	0.500	0.2500
0.841621	0.400	0.2000
1.000000	0.317	0.1585
1.036433	0.300	0.1500
1.281552	0.200	0.1000
1.644854	0.100	0.0500
1.959964	0.050	0.0250
2.326345	0.020	0.0100
2.575829	0.010	0.0050
2.807033	0.005	0.0025
3.090232	0.002	0.0010
3.290527	0.001	0.0005
3.480756	0.0005	0.00025
3.719016	0.0002	0.00010
3.890592	0.0001	0.00005
4.055630	0.00005	0.00025
4.264891	0.00002	0.000010
4.417173	0.00001	0.000005

Tabelle 4: Ausgewählte Schranken der Gauß-Verteilung

Die Funktion f(x) der Gauß-Verteilung wird auch als *Dichtefunktion* bezeichnet. Der Begriff "Dichte" ist eine Abkürzung für *Wahrscheinlichkeitsdichte* und bezieht sich auf die Bedeutung, die die Gauß-Verteilung nicht nur als mathematische Funktion, sondern auch als die vielleicht wichtigste statistische Wahrscheinlichkeitsverteilung besitzt. Trägt man in Abhängigkeit von x das Integral von $-\infty$ bis x ab, so erhält man die sogenannte *Verteilungsfunktion* $\Phi(x)$ der Gauß-Verteilung, die als "kumulative" Funktion einen typisch sigmoidförmigen Verlauf besitzt.

Die Fläche unter der Kurve lässt sich auch als *Wahrscheinlichkeit* interpretieren. Die Gesamtfläche unter der Kurve beträgt 1 (zweites Kolmogoroffsches Axiom!), und jedem Intervall kommt ein Anteil dieser Gesamtfläche zu. Dieser Gedanke ist zunächst etwas ungewohnt, da bisher nur mit diskreten Wahrscheinlichkeiten umgegangen wurde, wie dies beispielsweise beim Würfelspiel der Fall ist. Beim Würfeln ist bekannt, dass nur die Zahlen von 1 bis 6 auftreten können und dass jeder dieser Zahlen eine Wahrscheinlichkeit zukommt, mit der die Zufallsvariable "Augenzahl" eine konkrete Ausprägung, also eine der sechs diskreten Zahlen realisiert. Die Gauß-Verteilung dagegen ist eine stetige Funktion, mithin müssen alle reellen Zahlen als mögliche Realisationen (d.h. "Ergebnisse") in Frage kommen. Es macht demzufolge auch keinen Sinn mehr, nach der Wahrscheinlichkeit für die Zahl 1.234 oder für die Zahl π oder eine andere zu fragen, denn jede Zahl ist ja nur eine von unendlich vielen, die als Realisationen in Frage kommen und besitzt damit die Wahrscheinlichkeit 0.

Bei stetigen Skalen fragt man sich deshalb nicht nach der Wahrscheinlichkeit für *eine* Zahl, sondern nach der Wahrscheinlichkeit für ein *Intervall* auf der Abszisse: Wie groß ist die Wahrscheinlichkeit, dass – bei unterstellt stetiger Skala – ein in Frankfurt neugeborenes Kind zwischen 40 und 45 cm groß ist, oder höchstens 50 cm groß ist. Ähnlich wie man bei einem stetigen Histogramm einen Balken (d.h. eine Fläche) über einem Intervall proportional zur relativen Häufigkeit abträgt, kann man bei der stetigen Wahrscheinlichkeitsdichte den Flächenanteil zwischen zwei Abszissenpunkten als Maß für die Wahrscheinlichkeit verwenden, dass ein zufällig zu erhaltender Wert (ein Messwert im Sinne einer Realisation einer Gauß-verteilten Zufallsvariablen!) zwischen eben diesen beiden Abszissenwerten liegt. Also darf man die Überlegung, dass zum Beispiel zwischen $\mu\pm\sigma$ ein Flächenanteil von 68.3% liegt, auch "umdrehen": Mit einer Wahrscheinlichkeit von P=68.3%=0.683 wird ein zufällig gewonnener Wert zwischen den beiden Abszissenpunkten $\mu-\sigma$ und $\mu+\sigma$ liegen.

Zur Veranschaulichung kann man einen lebenspraktischen Versuch unternehmen. Man nimmt einen Eimer mit Sand, dessen Körnchen so klein sind, dass "beliebig viele" im Eimer Platz haben. Um den Versuch etwas spannender zu machen, wird eines der Sandkörnchen vorher grün angestrichen und der Sand gut durchmischt. Der Sand wird nun vorsichtig, ohne den Eimer zu verwackeln, auf einen Tisch geschüttet. Vom Profil her ergibt der Sandhaufen eine Linie, die mit der "Glockenkurve" der Gauß-Verteilung identifiziert werden kann; im zentralen Bereich liegen jedenfalls sehr viel mehr Körnchen als an den Rändern. Welcher Anteil der Körnchen liegt zwischen zwei Stellen, die mit $\mu-\sigma$ und $\mu+\sigma$ bezeichnet werden können? Man zählt die Körnchen in diesem Bereich und stellt vielleicht fest, dass gerade 68.3% zwischen diesen Grenzen liegen. Wie groß ist die Wahrscheinlichkeit, dass das grüne Körnchen zwischen $\mu\pm\sigma$ liegt? Offenbar P=0.683. Wie groß ist die Wahrscheinlichkeit, dass ein zufällig herausgegriffenes Körnchen zwischen $\mu-\sigma$ und $\mu+\sigma$ liegt? Da nun mal 68.3% aller Körnchen zwischen $\mu-\sigma$ und $\mu+\sigma$ liegen, ist diese Wahrscheinlichkeit wohl P=0.683, die gleiche wie für das grüne Körnchen. (Vorsicht: Die Sandkörnchen entsprechen Merkmalsträgern mit Werten in einem Intervall auf der Abszisse, nicht zu verwechseln mit der "Anzahl Zahlen" in diesem Intervall.)

Résumé: Betrachtet man eine Gauß-verteilte Zufallsvariable X, so wird eine zufällige Realisation X=x (ein zukünftig gemessener Wert also) mit der Wahrscheinlichkeit P zwischen den Grenzen $\mu-u_P\cdot\sigma$ und $\mu+u_P\cdot\sigma$ liegen, formal: $P = \text{Prob}(\mu-u_P\cdot\sigma \le X \le \mu+u_P\cdot\sigma)$.

Vermutlich findet man nirgends in der Medizin und Biologie exakte Gauß-Verteilungen vor, so dass eine Anwendung der Theorie fraglich sein könnte. Allenfalls wird man "einigermaßen" eingipflig und symmetrisch aussehende Verteilungen vorfinden, denen das mathematische Modell *Gauß-* bzw. *Normalverteilung* vielleicht "möglichst gut" nahe kommt. Glücklicherweise kann man aber diese strenge Betrachtung etwas abschwächen, denn man interessiert sich doch an dieser Stelle weniger für die Verteilung der Einzelwerte bzw. Messwerte x, sondern vielmehr für die Verteilung der Durchschnittswerte $\bar{x}$ und, insbesondere, für deren Variabilität bzw. "Genauigkeit". Eine der wenigen Ausnahmen davon wird im Kapitel 6.3 zum Thema "Normbereiche" angesprochen.

Eine Antwort auf die Frage nach der "Genauigkeit" der Durchschnittswerte $\bar{x}$ findet sich mit dem *Zentralen Grenzwertsatz* im nächsten Abschnitt:

4.3 Das Konfidenzintervall für den Erwartungswert μ

Untersucht man eine Gauß-verteilte Zufallsvariable X mit Erwartungswert ("Mittelwert") μ und der Varianz σ^2, so folgen die Durchschnittswerte $\bar{x}$ aus Stichproben zu je n Realisationen ("Messwerten") X=x ihrerseits einer Gauß-Verteilung. Die Durchschnittswerte $\bar{x}$ besitzen ebenfalls den Erwartungswert μ, jedoch die kleinere Varianz $\sigma_{\bar{x}}^2=\sigma^2/n$.

Ist die Verteilung der Ausgangswerte x keine exakte Gauß-Verteilung, so ist die Verteilung der Durchschnittswerte $\bar{x}$ trotzdem angenähert Gauß-verteilt: Diese Annäherung gilt umso besser, je größer der Stichprobenumfang n ist. Dieser Zusammenhang ist Aussage des sogenannten *Zentralen Grenzwertsatzes* der Statistik.

Der Quotient $\sigma_{\bar{x}}^2=\sigma^2/n$ ist plausibel, denn wenn n groß ist bzw. gegen unendlich strebt, dann wird $\bar{x}$ immer besser mit μ übereinstimmen bzw. gegen μ streben. Ist im anderen Grenzfall n=1, so ist $\sigma_{\bar{x}}^2=\sigma^2$ und $\bar{x}=x$, was ebenfalls plausibel ist.

Den Erwartungswert und die Varianz einer Funktion $F(x_1,...,x_n)$ von *zufälligen Größen* bezeichnet man auch mit den Symbolen E[F] und V[F]. Da $E[x_i]=\mu_i=\mu$, $V[x_i]=\sigma_i^2=\sigma^2$ und andererseits $E[\Sigma a_i \cdot x_i]=\Sigma(a_i \cdot E[x_i])$ und $V[\Sigma a_i \cdot x_i]=\Sigma(a_i^2 \cdot V[x_i])$ ist, ergibt sich eine mehr mathematische Begründung für die Beziehungen des Zentralen Grenzwertsatzes:

$$E[\bar{x}] = E[\frac{1}{n}\cdot\sum x_i] = \frac{1}{n}\cdot E[\sum x_i] = \frac{1}{n}\cdot\sum E[x_i] = \frac{1}{n}\cdot\sum \mu = \frac{1}{n}\cdot n\cdot\mu = \mu$$

$$V[\bar{x}] = V[\frac{1}{n}\cdot\sum x_i] = \frac{1}{n^2}\cdot V[\sum x_i] = \frac{1}{n^2}\cdot\sum V[x_i] = \frac{1}{n^2}\cdot\sum \sigma^2 = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{1}{n}\sigma^2$$

Für praktische Zwecke kann man daraus ableiten, dass eine Verteilung von Werten angenähert eingipflig und symmetrisch sein sollte, um davon ausgehen zu dürfen, dass die Verteilung der Durchschnittswerte wiederholter Stichproben in guter Näherung eine Gauß-Verteilung darstellt. Da in einer konkreten Untersuchung natürlich keine wiederholten Stichproben, sondern nur genau eine vorliegt, müsste die letzte Formulierung sicher adäquater lauten "dass der Durchschnitt $\bar{x}$ einer Gauß-Verteilung entstammt".

Eingipfligkeit ist – gegebenenfalls unter Zuhilfenahme der Überlegungen aus Abschnitt 5.2 – in der Regel relativ einfach zu beurteilen. Bei Verdacht auf Mehrgipfligkeit sollte immer eine Zerlegung der vielleicht vorliegenden

Mischpopulation versucht werden. Ein triviales Beispiel dazu: Umfasst eine Stichprobe Messwerte von Frauen und Männern, so kann dies - man denke nur an Hormonwerte oder an anthropometrische Daten - eine mehrgipflige Verteilung bedingen.

Die Symmetrie einer empirischen Verteilung zu beurteilen ist mitunter schwierig, so dass man einen statistischen Test zur Entscheidungshilfe heranziehen sollte (nächstes Kapitel, Abschnitt 5.2). In vielen Fällen kann man sogar medizinisch begründen, dass eine konkrete Verteilung keine Gauß-Verteilung sein kann: Dies ist praktisch immer bei Laborwerten der Fall, denn diese sind als Konzentrationen nach unten durch "0" begrenzt, die meisten Werte liegen in einem unteren Bereich, größere Werte kommen vor, sehr große Werte sind nicht ausgeschlossen. Die letzte Skizze führt zu einer typisch rechtsschiefen Verteilung, für die in Abbildung 16 mit dem Vergleich zweier Schilddrüsendiagnosen bezüglich TSH ein Beispiel angegeben wird. Einige Beispiele für offenbar symmetrische Verteilungen sind bereits bekannt, man denke nur an die Körpergrößen der Neugeborenen oder an die Verteilung der Leukozyten gesunder Probanden. Auf die Problematik schiefer Verteilungen wird im nächsten Abschnitt noch kurz eingegangen.

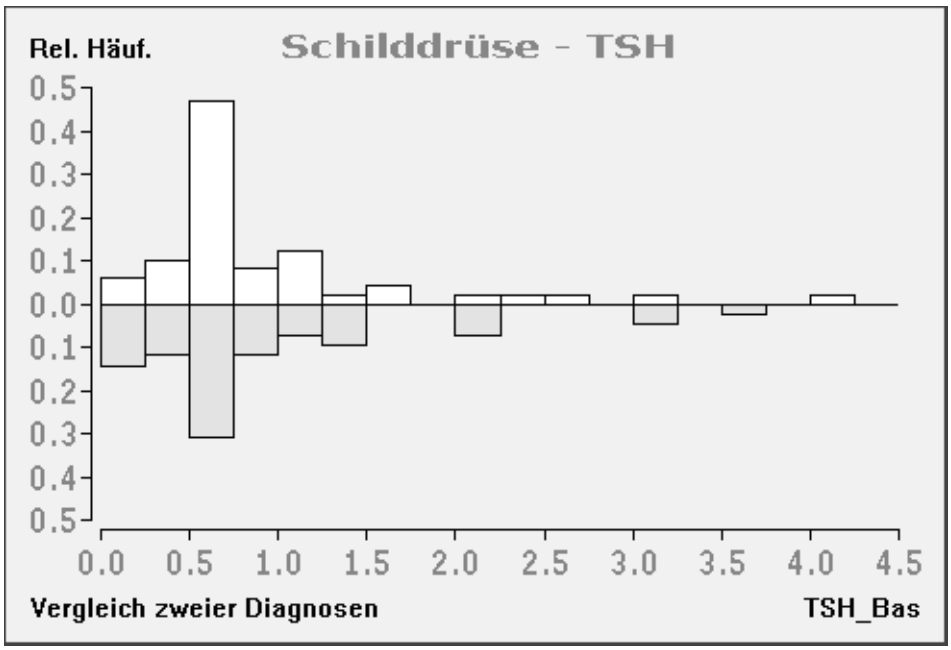


Abbildung 16: Beispiel für rechtsschiefe Häufigkeitsverteilungen

Ist die Voraussetzung "Eingipfligkeit und angenäherte Symmetrie der Verteilung der Messwerte" erfüllt, so gilt - man vergleiche Abschnitt 4.2 - die aus der Gauß-Verteilung abgeleitete Beziehung

$$\mu - u_P \cdot \sigma \leq x \leq \mu + u_P \cdot \sigma$$

angenähert mit der Wahrscheinlichkeit P. (Die Beziehung gilt exakt, wenn die Verteilung der Messwerte x eine exakte Gauß-Verteilung ist, bitte vergleichen Sie dazu den Zentralen Grenzwertsatz!) Da hier nicht von x, sondern von $\bar{x}$ die Rede ist, deshalb auch von der Varianz $\sigma_{\bar{x}}^2$ der Durchschnittswerte $\bar{x}$ und nicht von der Varianz σ^2 der Ausgangswerte x, ersetzt man x durch $\bar{x}$ und σ durch $\sigma_{\bar{x}}$ und erhält

$$\mu - u_P \cdot \sigma_{\bar{x}} \leq \bar{x} \leq \mu + u_P \cdot \sigma_{\bar{x}}$$

ebenfalls mit Wahrscheinlichkeit P. An Stelle von $\sigma_{\bar{x}}$ kann man (Zentraler Grenzwertsatz, 1. und 2. Absatz dieses Abschnittes!) natürlich auch $\sigma/\sqrt{n}$ einsetzen, so dass, ebenfalls mit Wahrscheinlichkeit P,

$$\mu - u_P \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + u_P \cdot \frac{\sigma}{\sqrt{n}}$$

ist. Die letzte Ungleichung beinhaltet vermöge μ und σ eine Aussage bezüglich $\bar{x}$, wünschenswert ist jedoch, umgekehrt mit Hilfe von $\bar{x}$ eine Aussage bezüglich μ zu treffen, denn letzteres ist ja unbekannt. Also formt man die Relation um und erhält durch Subtrahieren von $\bar{x}$ und μ in allen drei Teilen der Ungleichung

$$\bar{x} - u_P \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + u_P \cdot \frac{\sigma}{\sqrt{n}}$$

Wenn σ bekannt ist, kann man diese Ungleichung leicht interpretieren: Das letzte Intervall überdeckt den unbekannten Mittelwert μ der Grundgesamtheit mit einer Wahrscheinlichkeit von (vorgegeben!) P, zum Beispiel P=95%. Man nennt das Intervall deshalb auch *Konfidenzintervall*, wörtlich übersetzt auch *Vertrauensbereich* für μ.

Mit dem Konfidenzintervall hat man die in diesem Skriptum erste Aussage vorliegen, die mit einer Wahrscheinlichkeit (Konfidenz, "Sicherheit") P verbunden ist. Gleichzeitig ist auch das Ziel erreicht, nicht nur eine *Punktschätzung* für μ in Gestalt des Durchschnittes $\bar{x}$, sondern eine *Intervallschätzung* für μ zur Beurteilung der Genauigkeit der statistischen Schätzung anzugeben.

In praxi ist die Standardabweichung σ der Grundgesamtheit unbekannt, es ist aber naheliegend, die Größe σ durch ihre Schätzgröße s zu ersetzen. Diese besitzt jedoch keinen festen Wert, sondern unterliegt ihrerseits der Stichprobenvariabilität: Um nun trotz dieser Variabilität die vorgegebene Konfidenz P zu gewährleisten, wird u_P - hier ohne mathematische Begründung, vgl. dazu aber auch den nächster Absatz - durch eine korrespondierende Größe $t_{P,n-1}$ ersetzt, die ganz analog zu erklären ist wie der entsprechende Faktor u_P der Gauß-Verteilung. Die $t_{P,n-1}$ zugehörige Ver-

teilung wird üblicherweise als "t-Verteilung" bezeichnet und geht auf den englischen Mathematiker und Guinness-Mitarbeiter William Sealey Gosset (1876-1937) zurück, der diese Zusammenhänge im Jahre 1908 unter dem Pseudonym "Student" publizierte. Gosset's Größe $t_{P,n-1}$ hängt offensichtlich nicht nur von der Konfidenz P, sondern - zweiter Index! - auch vom Stichprobenumfang n ab; fg=n-1 bezeichnet man als *Freiheitsgrad* der t-Verteilung (und auch der Standardabweichung s, vgl. Abschnitt 3.2!). Für größere Werte von n (ab etwa $n>20$) stimmen u_P und $t_{P,n-1}$ recht gut überein, da die Variabilität von s für wachsendes n abnimmt. Je größer n ist, umso besser ist diese Übereinstimmung: Die t-Verteilung stimmt für $n\to\infty$ mit der Gauß-Verteilung überein.

Die Notwendigkeit der Ersetzung von u_P durch $t_{P,n-1}$ kann man sich leicht klar machen: Aus der Formel des Konfidenzintervalls ergibt sich, dass die Beziehung $u = |\bar{x}-\mu|/\sigma_{\bar{x}} \le u_P$ mit Wahrscheinlichkeit P zutrifft; dabei ist $\bar{x}$ die hier einzige Zufallsvariable, die - standardisiert via μ und $\sigma_{\bar{x}}$ - einer Gauß-Verteilung folgt. Ersetzt man jedoch in der letzten Beziehung $\sigma_{\bar{x}}$ durch $s_{\bar{x}}=s/\sqrt{n}$, so erhält man auf der linken Seite die Größe $t=|\bar{x}-\mu|/s_{\bar{x}}$, in die offenbar zwei errechnete Zufallsvariablen eingehen: Der Zähler als lineare Funktion von Gauß-verteilten Variablen ist Gauß-verteilt, im Nenner steht jetzt als weitere Zufallsvariable eine Funktion der Summe von quadrierten Gauß-verteilten Größen. Der Quotient t der beiden Funktionen wird somit vermutlich nicht - wie u - einer Gauß-Verteilung folgen: W.S. Gosset gelang es, die Wahrscheinlichkeitsverteilung von t und deren Percentilen $t_{P,n-1}$ zu ermittelt und konnte damit zeigen, dass nicht die Ungleichung $t = |\bar{x}-\mu|/s_{\bar{x}} \le u_P$, sondern nur die Beziehung $t = |\bar{x}-\mu|/s_{\bar{x}} \le t_{P,n-1}$ mit der Wahrscheinlichkeit P erfüllt ist. Damit muss aber offenbar auch Student's $t_{P,n-1}$ die Gaußsche Größe u_P im Konfidenzintervall ersetzen:

Nach den erwähnten Einsetzungen ergibt sich die endgültige Form des P·100%-Konfidenzintervalls für μ:

$$\bar{x} - t_{P,n-1} \cdot \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{P,n-1} \cdot \frac{s}{\sqrt{n}}$$

Dazu ein praktisches Beispiel: Denkt man wieder an die pädiatrische Untersuchung aus Kapitel 3, so ist dort die Berechnung eines Konfidenzintervalls sicher naheliegend; die Voraussetzungen sind erfüllt, denn Körpergrößen sind quantitativer Natur und die Verteilung der Daten kann als einigermaßen eingipflig und symmetrisch angenommen werden (Abbildungen 7 und 8). Der Wert $t_{P,n-1}=t_{0.95,70-1}=1.995$ stammt aus einer Tabelle der t-Verteilung (nächste Tabelle 5, interpoliert und gerundet) und weicht nur unwesentlich von dem Wert $u_{0.95}=1.96$ ab. Die Standardabweichung s und das arithmetische Mittel $\bar{x}$ wurden bereits oben berechnet (vgl. Abbildung 5), n=70 ist bekannt. Vermöge der letzten Formel ergibt sich daraus als Resultat das 95%-Konfidenzintervall mit der linken und rechten Grenze von 48.91[cm] bzw. 50.99[cm]: Man ist sich zu P=95% sicher, dass das Konfidenzintervall (48.91,50.99) [cm] die unbekannte, mittlere Körpergröße μ von Neugeborenen überdeckt. Damit trifft man eine Aussage, die weit über die oben errechnete Punktschätzung $\bar{x}=49.95$[cm] hinausgeht und, sogar mit einer Sicherheit P verbunden, eine Information zur Genauigkeit der Schätzung von μ beinhaltet.

Die häufig getroffene Aussage "μ liegt mit einer Sicherheit von 95% im Konfidenzintervall" ist nicht ganz korrekt: Grundsätzlich muss man sich den Parameter μ als einen zwar unbekannten, aber doch festen Wert vorstellen, das Konfidenzintervall als zufallsabhängige Größe dagegen variiert von Stichprobe zu Stichprobe. Stellt man sich μ als einen Pflock vor, der im Rasen steckt, und interpretiert einen Wurf mit einem Ring als Konfidenzintervallberechnung, so wird niemand behaupten, dass nach dem Wurf der Pflock plötzlich im Ring steckt!

Die letzte Zeile in Tabelle 5 für "Freiheitsgrad ∞" entspricht exakt den Werten in Tabelle 4 der Gauß-Verteilung. Dies ist nicht verwunderlich, denn oben wurde festgestellt, dass mit $n\to\infty$ auch $\overline{x}\to\mu$, $s^2\to\sigma^2$ und nicht zuletzt auch $t_{P,n-1}\to u_P$ strebt. So gesehen kann man auf die Tabelle 4 der Gauß-Verteilung verzichten.

Tabelle 5 ist nicht "vollständig", das heißt, es werden nur Werte von t für ausgewählte Freiheitsgrade angegeben. Falls ein gewünschter Freiheitsgrad nicht vorhanden ist, so kann man in diesem Fall eine lineare Interpolation vornehmen, um den gesuchten Wert von t zu erhalten; zum Beispiel erhält man damit den Tabellenwert von $t_{1\text{-}0.05,42}$ durch $t_{0.95,42}=t_{0.95,40}-2\cdot(t_{0.95,40}-t_{0.95,45})/5=2.021-2\cdot(2.021-2.014)/5=2.0182$. Ausführlichere Tabellen finden sich in dem bereits mehrfach erwähnten Buch von Lothar Sachs (Springer 2004/2018).

fg	Irrtumswahrscheinlichkeit α=1-P (zweiseitig)				
	0.0500	0.0200	0.0100	0.0010	0.0001
1	12.706	31.821	63.657	636.619	6366.198
2	4.303	6.965	9.925	31.598	99.992
3	3.182	4.541	5.841	12.924	28.000
4	2.776	3.747	4.604	8.610	15.544
5	2.571	3.365	4.032	6.869	11.178
6	2.447	3.143	3.707	5.959	9.082
7	2.365	2.998	3.499	5.408	7.885
8	2.306	2.896	3.355	5.041	7.120
9	2.262	2.821	3.250	4.781	6.594
10	2.228	2.764	3.169	4.587	6.211
11	2.201	2.718	3.106	4.437	5.921
12	2.179	2.681	3.055	4.318	5.694
13	2.160	2.650	3.012	4.221	5.512
14	2.145	2.625	2.977	4.140	5.363
15	2.131	2.602	2.947	4.073	5.239
16	2.120	2.583	2.921	4.015	5.134
17	2.110	2.567	2.898	3.965	5.044
18	2.101	2.552	2.878	3.922	4.965
19	2.093	2.539	2.861	3.883	4.898
20	2.086	2.528	2.845	3.850	4.837

Tabelle 5a: Tabelle der t-Verteilung für fg=1 bis fg=20

Bei Verwendung eines einschlägigen Computerprogramms - zum Beispiel **BiAS.** - benötigt man in der Regel keine Tabellen, da die erforderlichen Werte der t-Verteilung vom Programm berechnet werden. Die Umkehrung ist in den meisten Programmen ebenfalls vorgesehen: Zu einem gegebenen Wert von t und bekanntem Freiheitsgrad kann die zugehörige Irrtumswahrscheinlichkeit - im Kontext des nächsten Kapitels auch *Überschreitungswahrscheinlichkeit* oder *p-Wert* genannt - errechnet werden.

fg	Irrtumswahrscheinlichkeit α=1-P (zweiseitig)				
	0.0500	0.0200	0.0100	0.0010	0.0001
20	2.086	2.528	2.845	3.850	4.837
22	2.074	2.508	2.819	3.792	4.736
24	2.064	2.492	2.797	3.745	4.654
26	2.056	2.479	2.779	3.707	4.587
28	2.048	2.467	2.763	3.674	4.530
30	2.042	2.457	2.750	3.646	4.482
32	2.037	2.449	2.738	3.622	4.441
34	2.032	2.441	2.728	3.601	4.405
36	2.028	2.434	2.719	3.582	4.373
38	2.024	2.429	2.712	3.566	4.346
40	2.021	2.423	2.704	3.551	4.321
45	2.014	2.412	2.690	3.520	4.269
50	2.009	2.403	2.678	3.496	4.228
55	2.004	2.396	2.668	3.476	4.196
60	2.000	2.390	2.660	3.460	4.169
65	1.997	2.385	2.654	3.447	4.146
70	1.994	2.381	2.648	3.435	4.127
75	1.992	2.377	2.643	3.425	4.110
80	1.990	2.374	2.639	3.416	4.096
85	1.988	2.371	2.635	3.409	4.083
90	1.987	2.368	2.632	3.402	4.072
95	1.985	2.366	2.629	3.396	4.062
100	1.984	2.364	2.626	3.391	4.053
200	1.972	2.345	2.601	3.340	3.970
300	1.968	2.339	2.592	3.323	3.943
400	1.966	2.336	2.588	3.315	3.930
500	1.965	2.334	2.586	3.310	3.922
600	1.964	2.333	2.584	3.307	3.917
700	1.963	2.332	2.583	3.305	3.913
800	1.963	2.331	2.582	3.303	3.910
900	1.963	2.330	2.581	3.301	3.908
1000	1.962	2.330	2.581	3.300	3.906
∞	1.960	2.326	2.576	3.290	3.891

Tabelle 5b: Tabelle der t-Verteilung für fg=20 bis fg=1000

Falls eine der Voraussetzungen zur Berechnung eines Konfidenzintervalls nicht erfüllt ist, die Daten also nicht-quantitativ sind und/oder die Verteilung schief ist, so kommt alternativ ein nicht-parametrisches Konfidenzintervall für den Median in Frage. (Nicht-parametrische Methoden erfordern keine Gauß-Verteilungen. Das oben besprochene Konfidenzintervall ist ein parametrisches Intervall.) Dazu ordnet man wieder die Stichprobe der Größe nach und

erhält $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$. Für Stichprobenumfänge $n \geq 50$ errechnet man eine Hilfsgröße h mit $h=(n-1-u_P \cdot \sqrt{n})/2$, wobei P die gewünschte Konfidenz und u_P der zugehörige Wert der standardisierten Gauß-Verteilung ist (Tabelle 4, z.B. ist $u_{0.95}=1.96$). Das nicht-parametrische Konfidenzintervall für den Median der Grundgesamtheit errechnet sich mit Hilfe der Beziehung

$$x_{(h)} \leq \tilde{\mu} \leq x_{(n-h+1)} \text{ mit } h=(n-u_P \cdot \sqrt{n}-1)/2$$

Für kleinere Stichprobenumfänge n<50 muss man Tabellen verwenden. Hinweise auf die mathematische Herleitung und Tabellen finden sich bei Sachs (2004/2018).

4.4 Wahl der Konfidenz P

Offen ist noch die Festlegung der Konfidenz P, den Wert μ zu überdecken bzw. des Risikos $\alpha=1-P$, den Parameter μ nicht zu überdecken. Aus der Formel $\bar{x} \pm t_{P,n-1} \cdot s/\sqrt{n}$ für das Konfidenzintervall ist ersichtlich, dass das Intervall "klein" wird, wenn der Stichprobenumfang n, der im Nenner steht, "groß" ist (das Umgekehrte gilt natürlich auch) und andererseits das Intervall "groß" wird, wenn P "groß" ist (siehe t-Tabelle, die Umkehrung gilt auch hier!). Da ein kleiner Stichprobenumfang n also ein relativ breites Konfidenzintervall bewirkt, sollte man für kleine n zur Kompensation eine eher "kleine" Konfidenz P wählen: Dies ist die Situation eines Pilot-Versuches, der bei geringem Aufwand, d.h. bei kleiner Fallzahl n ein noch relativ großes Risiko von $\alpha=1-P=1-0.95=0.05$ impliziert, den Parameter μ nicht zu überdecken. Als zweite Kombination ist ein großes n bei großer Konfidenz P denkbar: Dies ist die Situation einer bestätigenden Studie, in der ein bereits bekannter Sachverhalt zum Beispiel zwecks behördlicher Zulassung (BfArM) eines Medikamentes noch einmal gesichert werden soll: Zur Absicherung eines vermuteten bzw. prinzipiell bekannten Sachverhaltes ist eine große Fallzahl n erforderlich, die wiederum eine große Konfidenz P zulässt.

Pilot-Studien sind den Phasen I und II, bestätigende Studien der Phase III zuzuordnen (zur Klassifikation vgl. Abschnitt 1.4). Andere Fragestellungen, die nicht unbedingt mit Medikamentenforschung zu tun haben, sind natürlich gleichermaßen denkbar: Man stelle sich vor, dass eben ein neues Isoenzym der Alkalischen Phosphatase entdeckt wurde, dessen Bestimmung im Labor aber noch recht teuer und vielleicht zeitaufwendig ist. Selbstredend wird man versuchen, mit relativ wenigen Bestimmungen einen Eindruck über die Größenordnung der möglichen Messwerte zu erhalten und muss in diesem Kontext auch bereit sein, eine Konfidenz P von "nur" 95% zu akzeptieren.

Konventionellerweise verwendet man in Abhängigkeit von der medizinischen Fragestellung Konfidenzen P von 0.95, 0.99 und gelegentlich sogar 0.999. Wann aber ist nun ein Stichprobenumfang n als "groß" oder "klein" zu betrachten?

4.5 Fallzahlberechnungen

Um eine sachgerechte Fallzahlabschätzung für ein geplantes Konfidenzintervall zu erhalten, muss man aus früheren Studien oder aus einer Pilot-Studie gewisse Vorstellungen über die Variabilität (im Sinne der Streuung s^2 bzw. der Varianz σ^2) der erwarteten Messwerte einbringen, ferner muss neben einer Festlegung der Konfidenz P entschieden werden, wie lang das Konfidenzintervall aus medizinischer Sicht zweckmäßigerweise höchstens sein sollte; diese Länge sei hier mit dem Symbol "L" bezeichnet. Aus der Formel des parametrischen Konfidenzintervalls bei bekannter Standardabweichung σ (vgl. Abschnitt 4.3) ergibt sich somit

$$L = 2 \cdot u_P \cdot \frac{\sigma}{\sqrt{n}}$$

und daraus wiederum

$$n = (2 \cdot u_P \cdot \sigma / L)^2$$

Bei unbekannter Standardabweichung σ setzt man - zum Beispiel aus Voruntersuchungen oder aus der Literatur - als Schätzung die empirische Standardabweichung s ein und erhält auf diesem Wege eine Schätzung für die erforderliche Fallzahl.

Für P=0.95 kann man eine einfache "Faustregel" ableiten: Für P=0.95 ist $u_P=u_{0.95}=1.96\approx 2$, dies in die letzte Formel eingesetzt ergibt

$$n \approx (4 \cdot \sigma / L)^2$$

Noch einfacher wird es, wenn man die Standardabweichung s bzw. σ mit Hilfe der geschätzten, vermuteten oder bekannten Spannweite R abschätzt: Für mittlere Stichprobenumfänge n ist $R\approx 4s$. Damit reduziert sich die Fallzahlberechnung auf die folgende "Faustregel":

$$n \approx (R / L)^2$$

Die hier abgeleiteten Formeln zur Berechnung der Fallzahl gelten streng genommen nur bei bekannter, nicht aber bei unbekannter Varianz σ^2 und unterschätzen - speziell bei kleinen Fallzahlen - systematisch den tatsächlich erforderlichen Stichprobenumfang. In der Praxis verwendet man zur Fallzahlberechnung in jedem Fall ein einschlägiges Computerprogramm, zum Beispiel "**BiAS.** für Windows".

Unabhängig davon kann man feststellen, dass es zu einer gegebenen Fragestellung offenbar keinen "großen" oder "kleinen", sondern - in Abhängigkeit vom medizinischen Kontext - nur einen "richtigen", im oben

verwendeten Sinne "sachgerechten" Stichprobenumfang gibt. Wenn das Konfidenzintervall nicht nur zur Intervallschätzung, sondern zur Prüfung von Nullhypothesen verwendet werden soll (dazu mehr im nächsten Kapitel!), so muss neben dem Risiko α=1-P für den Fehler 1. Art noch ein weiteres Risiko β für den sogenannten Fehler 2. Art (das in diesem Abschnitt implizit mit β=0.5 angenommen wird!) in die Überlegung eingehen, wodurch die Fallzahlberechnung etwas komplizierter wird. Abschnitt 5.10 gibt darüber näheren Aufschluss.

4.6 Ausblick

Konfidenzintervalle kann man in vielen Situationen und für grundsätzlich alle, also nicht nur für quantitative Daten ermitteln. Die beiden im vorletzten Abschnitt beschriebenen Konfidenzintervalle für eingipflig-symmetrisch verteilte, quantitative Daten können noch durch viele weitere, bei anderen Fragestellungen adäquate Intervalle ergänzt werden (eine interessante Variante findet sich z.B. in Abschnitt 5.7 mit Abbildung 22), was aber nicht in der Absicht dieses einführenden Skriptums liegt. Umfangreiche Lehrbücher und Nachschlagewerke wie zum Beispiel das von Lothar Sachs (Springer-Verlag 2004/18) bieten reichlich Gelegenheit zur weiterführenden Lektüre.

Der Autor ist der Hoffnung, dass die Intention und die allgemeine Bedeutung von Konfidenzintervallen - und dies nicht nur in den hier diskutierten Fällen! - deutlich geworden sind und, insbesondere, dass ausreichend darauf hingewiesen wurde, dass Konfidenzintervalle, falls nur irgend möglich, fester Bestandteil einer deskriptiven statistischen Auswertung sein sollten: In zahlreichen Fachzeitschriften wird immer wieder (zum Beispiel bei D.G. Altman, Statistical Reviewing for Medical Journals, Stat. in Med. 17, 1998, pp. 2661ff) auf diesen obligatorischen Beitrag hingewiesen.

Die GCP- bzw. ICH-Guidelines ("Statistical Priciples for Clinical Trials" Topic E9, 02/1998), damit auch referierte wissenschaftliche Fachzeitschriften und die Zulassungsbehörden (zum Beispiel BfArM, FDA etc.) verlangen *obligat* zu allen relevanten Parametern die Angabe von Konfidenzintervallen.

Mit Konfidenzintervallen kann man noch mehr als nur eine deskriptive Auswertung von Daten vornehmen. Das nächste Kapitel ist der statistischen Testtheorie - der sogenannten konfirmatorischen Statistik – gewidmet, die ihrerseits reichhaltigen Gebrauch von Konfidenzintervallen macht, in diesem Fall jedoch als Hilfsmittel zur Nullhypothesenprüfung.

Damit erreicht man über die Brücke der Konfidenzintervalle das Gebiet der schließenden Statistik:

Kapitel 5:
Statistische Testverfahren

Viele wissenschaftliche Fragestellungen kann man in der Praxis optimal und erschöpfend mit Hilfe von deskriptiven statistischen Methoden beantworten. Die Frage "Wie groß sind in Frankfurt geborene Kinder?" ist, wie beschrieben, am besten mit dem Durchschnitt $\bar{x}$, dem Konfidenzintervall für μ und vielleicht noch mit "Normbereichen" (Abschnitt 6.3) zu beantworten. Im Gegensatz dazu gibt es aber auch viele Probleme, die man auf diese Weise noch nicht ausreichend bearbeitet hat: Stimmen bei der Eichung eines Messgerätes die Messergebnisse mit den vorgegebenen Angaben überein? Zeigt ein Medikament eine Wirkung? Unterscheiden sich gewisse Blutspiegel bei verschiedenen Diagnosen? Soll in Zukunft die neue Therapie eingesetzt werden, oder bleibt man besser bei der konventionellen? Die letzten Fragen verlangen zweifelsohne eine eindeutige Antwort: Ja oder Nein. Mit den bisher besprochenen deskriptiven Methoden kann man dieses Ziel offenbar nicht erreichen, sondern man muss sich entscheidungsorientierter statistischer Methoden bedienen. Die Struktur solcher Methoden wird eingehend Gegenstand dieses Kapitels sein.

Die Methodenlehre der *Induktiven* (schließenden) oder *Konfirmatorischen Statistik* - unterteilt in *Parametrische* und *Nicht-parametrische Verfahren*, erstere setzen in aller Regel eine Gauß-Verteilung der Messwerte voraus, letztere dagegen nicht - ist sehr umfangreich, und die Aussicht auf eine einigermaßen vollständige Lektüre stellt vielleicht keine allzu reizvolle Perspektive dar. Ziel dieses Skriptums kann also nur sein, ein grundlegendes Verständnis für die statistische Methodik zu entwickeln: In diesem Kapitel werden deshalb für alle drei relevanten Skalentypen (also für quantitative, ordinale und nominale Skalen) exemplarisch drei typische Vertreter statistischer Testverfahren etwas eingehender beschrieben, um ein prinzipielles Verständnis und einen Zugang zu der Arbeitsweise mit allen Skalenvarianten zu vermitteln. Wie schon erwähnt, soll dabei möglichst parallel ein mehr intuitiver und ein mehr mathematisch orientierter Weg eingeschlagen werden; wer an einzelnen mathematischen Aspekten vielleicht weniger interessiert ist, mag sich die entsprechenden Passagen eher aus der Distanz betrachten, trotzdem aber stets für sich die drei Fragen beantworten: Zur Umsetzung der medizinischen Fragestellung in den statistischen Kontext - Wie lautet die Nullhypothese? Zur Methodenauswahl - Welche Voraussetzungen (Skalen etc.!) sind zu beachten? Welche Methoden kommen damit zur Nullhypothesenprüfung in Frage? Und zur Umsetzung des statistischen Ergebnisses in den medizinischen Kontext - Wie muss man das Ergebnis formulieren?

In Abschnitt 5.10 findet sich ein kurzer Ausblick auf weitere, in diesem Skript nicht behandelte Methoden der induktiven Statistik.

5.1 Das Testen von Nullhypothesen

Ein Kardiologe möchte in einem Pilotversuch die Wirkung einer neuen Substanz zur Behandlung der Hypertonie untersuchen. Seine medizinische Hypothese lautet "Die neue Substanz senkt den Blutdruck". Auf eine Kontrollgruppe will er vorläufig noch verzichten (cave: siehe Kapitel 2 zur Versuchsplanung und den Abschnitt 5.4 weiter unten!) und behandelt n=9 Patienten mit seinem neuen Medikament. Er erhält folgende Ergebnisse:

Systolischer Blutdruck unter der neuen Behandlung „Novum“		
vor Behandlung	**nach Behandlung**	**Differenz**
185	170	15
190	165	25
170	160	10
175	180	-5
190	185	5
180	180	0
185	165	20
180	170	10
180	175	5

Tabelle 6a: Ergebnisse der Hypertonie-Studie für „Novum“

Eine Auswertung des Versuchs erbrachte die in Abbildung 17a aufgeführten, zum Teil noch nicht verständlichen Resultate, die aber später in diesem und auch in Abschnitt 5.3 zum "Einstichproben-t-Test" zur Interpretation der Versuchsergebnisse bedeutsam sind:

Einstichproben-t-Test

```
Schätzwerte:
Xquer = 9.4    s = 9.5    s² = 90.3

Fallzahl: n = 9

Test auf Gauß-Verteilung der Differenzen:
Kolmogoroff-Smirnoff's δ = 0.1434  ( p > 0.20, Ok! )

Zweiseitiger Test der Nullhypothese Ho(μdiff=0):
Student's Prüfgröße t = 2.98 mit df=8  ( p = 0.017547 )

Konfidenzintervall für μ:
P = 0.95:  [ 2.1410 , 16.7479 ]
```

Abbildung 17a: Programmausgabe zur Hypertonie-Studie für „Novum“

Die deskriptiven Maße $\bar{x}$, s etc. sind dem Untersucher geläufig. Der Test auf Gauß-Verteilung ist ihm derzeit noch unklar und er beschließt, dazu einen Biometriker zu konsultieren (nächster Abschnitt 5.2). Als ersten Schritt hat er sich ein Diagramm seiner quantitativen (!) Differenzen angefertigt und findet seiner Einschätzung nach keine Evidenz gegen die Annahme einer Gauß-Verteilung, da die Werte nicht gerade unsymmetrisch liegen und auch keine augenfälligen "Ausreißer" zu erkennen sind. Wie aber soll der Kardiologe mit seiner Arbeitshypothese "Die Substanz senkt den Blutdruck" umgehen?

Da sich die Auswertung der Daten auf den Vergleich vor und nach Behandlung, also auf die Differenzen "vor-nach Therapie" stützt, beurteilt der Kardiologe - nach Überprüfung der Annahme einer Gauß-Verteilung - das 95%-Konfidenzintervall für die Wirkungsdifferenz "*vor-nach*" und stellt fest, dass er doch wohl mit einer Sicherheit von 95% davon ausgehen kann, dass die mittlere Blutdruckveränderung μ in der Grundgesamtheit aller vergleichbaren Hypertonie-Patienten zwischen 2.1 und 16.7 mmHg liegt. (Diese Aussage muss später nur noch geringfügig korrigiert werden.)

Der Kardiologe folgert, dass seine Substanz X mit einer Sicherheit von 95% den Blutdruck um mindestens 2.1 mmHg (untere Grenze des Konfidenzintervalls!) senkt und fasst dies als einen Beweis für seine Vermutung auf. (Auch diese Aussage wird nach der späteren Diskussion in diesem Abschnitt noch geringfügig zu korrigieren sein.)

Wie aber geht nun die Statistik das Problem an?

In der Statistik setzt man die medizinische Fragestellung zunächst in eine sogenannte *Nullhypothese* um. Die Nullhypothese lautet hier "Das Medikament zeigt keine Wirkung" (formal: $H_0(\mu=0)$, "Die mittlere Blutdrucksenkung μ in der Population aller vergleichbaren Patienten ist gleich Null"). Unter dieser Annahme wäre der in der Stichprobe beobachtete Effekt ein Werk des Zufalls. Abbildung 18 hilft bei der weiteren Entscheidung:

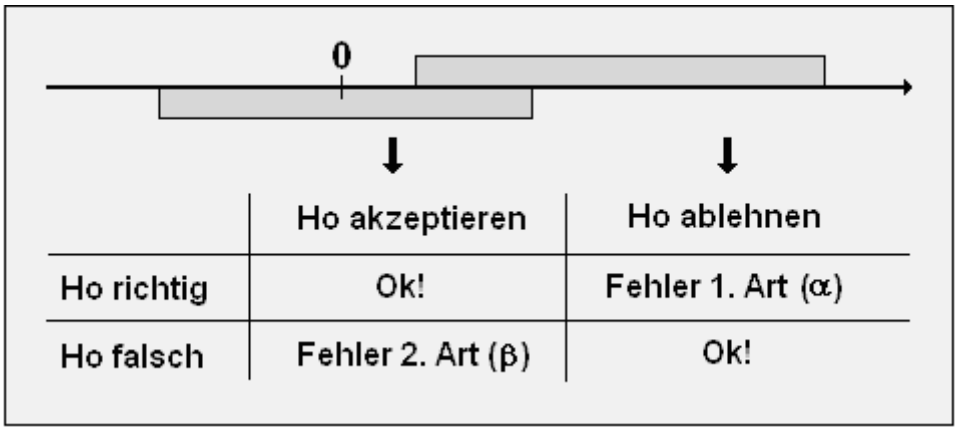


Abbildung 18: Entscheidungsregel zu Nullhypothesenprüfung

Angenommen, man erhält als Versuchsergebnis das linke Konfidenzintervall; dieses Konfidenzintervall überdeckt den Erwartungswert μ mit der gewählten Wahrscheinlichkeit P=1-α. Die Nullhypothese H_0 behauptet,

dass $\mu=0$ ist: Der Wert "Null" ist im Konfidenzintervall enthalten, das Versuchsergebnis ist mit der Annahme der H_0 verträglich, und es besteht somit kaum ein Grund, an der Aussage der Nullhypothese zu zweifeln. Tritt jedoch der zweite Fall des rechten Konfidenzintervalls ein, so ist ein Widerspruch zwischen der Aussage des Konfidenzintervalls und der Aussage der Nullhypothese festzustellen: Da einerseits das Konfidenzintervall mit Wahrscheinlichkeit P den unbekannten Wert μ überdeckt, andererseits aber die von H_0 postulierte "Null" außerhalb des Intervalls liegt, so ist diese Situation eher als Hinweis darauf aufzufassen, dass die Nullhypothese - und es ist ja nur eine Hypothese! - in Wirklichkeit falsch ist (und das neue Antihypertensivum also tatsächlich wirksam ist). Damit fällt die Entscheidung für die *Alternativhypothese* $H_A(\mu\neq0)$.

Bewiesen wurde die Nullhypothese im ersten Fall natürlich nicht, denn einerseits liegen ja noch andere Werte im Konfidenzintervall und andererseits wurde nur mit einer Wahrscheinlichkeit von $P<1$ (hier $P=0.95$) gearbeitet. Im Fall der Ablehnung der Nullhypothese gilt entsprechendes, denn hier geht man ja immerhin ein Risiko von $\alpha=1-P$ (hier $\alpha=0.05$) ein, μ nicht zu überdecken. Dies ist Gegenstand des zweiten Teils der Entscheidungstabelle: Ob H_0 definitiv richtig oder definitiv falsch ist, ist unbekannt. Womit ist in dem einen und in dem anderen Fall zu rechnen?

Angenommen, H_0 ist richtig: μ ist tatsächlich gleich Null, das Medikament des Kardiologen ist wirkungslos. Im Feld links oben (H_0 ist richtig, und wird auch via Konfidenzintervall akzeptiert) trifft man eine korrekte Entscheidung, im Feld rechts oben (H_0 ablehnen, H_0 ist aber per Annahme richtig!) dagegen eine Fehlentscheidung. Das Risiko für diesen sogenannten *Fehler 1. Art* ist einfach anzugeben: μ wird mit Wahrscheinlichkeit P überdeckt bzw. mit dem Risiko $\alpha=1-P$ nicht überdeckt. Ist also die Nullhypothese wahr und μ somit tatsächlich gleich Null, so geht man offenbar das Risiko α ein, μ (=0!) nicht zu überdecken und somit - in Unkenntnis der wahren Lage - zu folgern, dass H_0 falsch ist. Also: Mit dem Risiko α für den Fehler 1. Art wird H_0 abgelehnt.

Angenommen, H_0 ist falsch: μ ist ungleich Null, das Medikament des Kardiologen ist also wirksam. Über die richtige Entscheidung, H_0 abzulehnen und H_0 ist auch tatsächlich falsch, muss sicher nicht weiter gesprochen werden. Interessanter ist die Situation des *Fehlers 2. Art*, H_0 zu akzeptieren, obwohl H_0 falsch ist: Diesem Fehler 2. Art ist ein Risiko β zugeordnet, dessen Größe jedoch nicht allgemein angegeben werden kann. Im Beispiel des Kardiologen: Man schließt nach der in Abbildung 18 definierten Entscheidungsregel, dass das Medikament wirkungslos ist (H_0 wird akzeptiert!), obwohl vielleicht eine gewisse Wirkung doch vorhanden ist, nur: diese kann man vermöge des Konfidenzintervalls nicht aufzeigen, weil sie "zu klein" ist. Ist μ fast Null, so wird der hypothetische Wert "Null" fast immer dann überdeckt, wenn auch μ überdeckt wird. Dies geschieht per definitionem in fast $P\cdot100\%$ (im Beispiel: $P=95\%$) aller Fälle, und somit wird man H_0 mit einem Risiko β von nahezu P akzeptieren, obwohl

die Hypothese in Wahrheit falsch ist. Lässt man gedanklich μ von 0 wegrücken und allmählich größer werden, so wird sich auch das Risiko β für den Fehler 2. Art allmählich verkleinern, und, im Grenzfall für $\mu \to \infty$, verschwindend klein werden: Besitzt das neue Medikament des Kardiologen keine minimale, sondern eine recht ordentliche oder gar überzeugende Wirkung, so wird das Risiko, diese Wirkung per Nullhypothesenprüfung nicht zu erkennen, gering sein. Da aber nun a priori μ unbekannt ist, kann man auch keinen numerischen Wert für das Risiko β angeben: dieser kann nur bei Kenntnis des numerischen Wertes von μ bekannt sein.

Mit statistischen Testmethoden kann man also keine Beweise im Sinne der Mathematik führen, sondern man kann "nur" Entscheidungen unter Risiko treffen: Dies jedoch mit dem eindeutigen Vorteil einer quantitativen Vorgabe des objektiven Risikos α, denn damit gelangt jeder Untersucher zum gleichen Ergebnis, und jeder Leser wird dieses Ergebnis in gleicher Weise verstehen und interpretieren. In diesem Sinne ist damit das eingangs formulierte Ziel erreicht, mit objektiven Methoden objektive Entscheidungen zu treffen.

Eine statistisch-methodisch ganz andere, von der Entscheidungslogik her aber identische Fragestellung wurde bereits in Abschnitt 0.5 (Binomialverteilung) angesprochen: Die Inzidenzrate akuter Leukämien bei $\leq$4-jährigen wird in der BRD mit θ=0.000104 angegeben (Quelle: Keller et al. (1990), Untersuchung von Krebserkrankungen im Kindesalter in der Umgebung westdeutscher kerntechnischer Anlagen, Bundesministerium UNR, vgl. dazu auch Kaatsch et al. (2008) im Deutschen Ärzteblatt). Es wurde angenommen, dass in einem Umkreis von 30km um ein Atomkraftwerk n=4000 Kinder im Alter von bis zu 4 Jahren leben, unter denen innerhalb eines Jahres - bei erwarteten 0.42 Fällen - k=3 neue Fälle von akuter Leukämie erfasst werden (außer θ fiktive Daten). Mit Hilfe der Binomialverteilung wurde die Wahrscheinlichkeit $B_{\theta,4000,3}$=0.0088 für das Auftreten von drei oder mehr Fällen berechnet.

Als Nullhypothese formuliert man $H_O(\theta_{AKW}=\theta=0.000104)$. Unter dieser Nullhypothese ist $B_{\theta,4000,3}=0.0088<\alpha=0.01$, so dass die Nullhypothese an der Signifikanzschwelle α=0.01 zurückgewiesen werden kann. Die (hier fiktiven) Daten lassen daher mit einer Irrtumswahrscheinlichkeit von α=0.01 darauf schließen, dass die Leukämierate θ_{AKW} im Umkreis des AKW höher liegt als die für die BRD angegebene Rate von θ=0.000104.

Es ist zu beachten, dass hier eine *einseitige* Prüfung vorgenommen wurde, da aus biologischen Gründen eine Erniedrigung der Leukämierate ausgeschlossen werden kann; die einseitige Alternativhypothese lautet somit $H_A(\theta_{AKW}>\theta=0.000104)$ (und nicht $H_A(\theta_{AKW}\neq\theta)$!). In Abschnitt 6.1 (Qualitätssicherung) findet sich ein Beispiel für eine *zweiseitige* Fragestellung: Untersucht man einen möglichen zeitlichen Trend im Verlauf der Kontrollmessungen eines Laborgeräts, so kann man - mit θ=0.5 für eine Vergrößerung bzw. Verkleinerung der sukzessiven Kontrollmessungen - mit Hilfe der in Abschnitt 0.5 abgeleiteten Binomialverteilung bezüglich der H_O(„kein Trend“) einen *zweiseitigen* Test konstruieren.

Der am Beispiel der Leukämie-Inzidenzen abgeleitete Test wird in der Literatur als *Binomial-Test*, in anderen Zusammenhängen auch als *Vorzeichen-Test* bezeichnet. Interessierte Leserinnen und Leser finden in dem Buch von Lothar Sachs (2004/2018) weitere Einzelheiten und viele interessante Beispiele.

Das Prinzip der Nullhypothesenprüfung gemäß der Entscheidungsregel in Abbildung 18 ist für sämtliche statistischen Testverfahren gleich, so dass in den folgenden Abschnitten an zahlreichen Stellen darauf zurückgegriffen wird.

5.2 Test auf Gauß-Verteilung

Parametrische Verfahren wie das Konfidenzintervall oder die im Anschluss vorgestellten t-Tests setzen im mathematischen Modell voraus, dass die untersuchten Stichproben einer Gauß-Verteilung entstammen. Eine graphische, subjektive Überprüfung dieser Voraussetzung anhand eines Histogramms - wie in Abschnitt 5.1 - ist sicher hilfreich, trotzdem ist es wünschenswert, nicht nur "per Aspekt", sondern anhand einer objektiven Methode ein Urteil zu treffen.

Zur Prüfung der Nullhypothese *H_0(Die beobachteten Daten entstammen einer Gauß-Verteilung)* finden sich in der Literatur mehrere Verfahren, von denen hier der *Test von Kolmogoroff und Smirnoff* vorgestellt wird.

Vielfach wird der *Test von Shapiro und Wilk* als der "beste" Test auf Gauß-Verteilung bezeichnet. Der Kolmogoroff-Smirnoff-Test ist jedoch mathematisch besser abgesichert und auch anschaulich verständlich, so dass dessen Darstellung bevorzugt wurde. In praxi verwendet man eine Modifikation nach Lilliefors, Dallal und Wilkinson, dazu Sachs (2004/2018).

Abbildung 19 illustriert das Prinzip des Kolmogoroff-Smirnoff-Tests anhand der Daten des Hypertonie-Beispiels aus Abschnitt 5.1 (Daten der neuen Therapie). Die treppenförmige Kurve stellt die relative Häufigkeitssumme H_j (vgl. dazu Abschnitt 3.5) der n=9 Blutdruckdifferenzen dar. Die kontinuierliche Kurve ist die per H_0 erwartete (kumulative) Verteilungsfunktion $\phi(x)$ der Gauß-Verteilung, die anhand von $\bar{x}$ und s^2 als Schätzwerte für μ und σ^2 bestimmt wurde (zu den Definitionen vgl. Abschnitt 4.2).

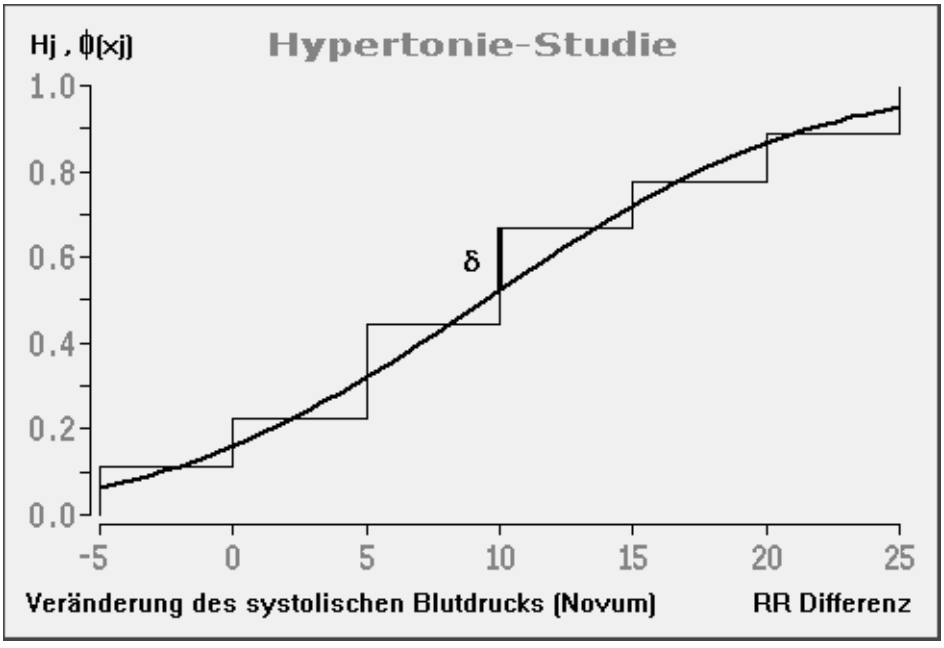


Abbildung 19: Illustration zum Kolmogoroff-Smirnoff-Test

Die größte vertikale Abweichung der empirischen (Stichprobe!) von der theoretischen Kurve der Gaußschen Verteilungsfunktion findet sich in Abbildung 19 bei einer Blutdruckdifferenz von 10 mmHg. Diese maximale Abweichung von δ=0.1434 entspricht gerade der Prüfgröße δ des Kolmogoroff-Smirnoff-Tests:

Bezeichnet man wieder die relative Häufigkeitssumme mit H_j und die unter der Nullhypothese "Gauß-Verteilung" erwarteten Anteile mit $\phi(x_j)$, so berechnet man die Prüfgröße δ mit

$$\delta = \max_{j=1}^{n} | H_j - \phi(x_j) |$$

und vergleicht diesen Wert mit einer vom vorgegebenen Signifikanzniveau α und vom Stichprobenumfang n abhängigen kritischen Schwelle $D_{\alpha,n}$: Überschreitet die berechnete Differenz δ die kritische Schwelle $D_{\alpha,n}$, so lehnt man die Nullhypothese ab und geht - mit einer Irrtumswahrscheinlichkeit von α, üblicherweise α=0.05 - davon aus, dass die Stichprobe *nicht* einer Gauß-Verteilung entstammt.

Eine mathematische Herleitung dieser Schwellenwerte ist sehr aufwendig, so dass man einfacher vorhandene Tabellen verwendet (zum Beispiel aus Sachs (2004/2018)). Dazu ein Auszug:

n	**5**	**8**	**10**	**12**	**15**	**20**	**25**	**30**
α=0.10	0.319	0.265	0.241	0.222	0.201	0.176	0.159	0.146
α=0.05	0.343	0.288	0.262	0.242	0.219	0.192	0.173	0.159

Tabelle 7: Kritische Werte $D_{\alpha,n}$ zum Kolmogoroff-Smirnoff-Test

Da die im Beispiel berechnete Differenz δ=0.1434 den kritischen Wert für n=9 sogar an der Signifikanzschwelle α=0.10 deutlich unterschreitet (interpoliert ist $D_{0.10,9}$=0.253), kann auf diesem Wege die Nullhypothese am Signifikanzniveau α=0.10 beibehalten werden, was die Konfidenzintervall-Berechnung des Kardiologen bzgl. der Voraussetzungen rechtfertigt.

Statistische Programme - so auch das Programmpaket **BiAS.** - berechnen an Stelle von Prüfgrößen und einem anschließenden Vergleich mit Tabellenwerten vielfach sogenannte p-Werte, die das minimale Signifikanzniveau α angeben, an dem die fragliche Nullhypothese gerade noch abgelehnt werden könnte:

Im letzten Beispiel ist dieser p-Wert größer als 0.20, womit die Nullhypothese akzeptiert wird. Wie bereits erwähnt, wurde der Kolmogoroff-Smirnoff-Test von einigen weiteren Autoren modifiziert, worauf hier nicht eingegangen wird; Hinweise dazu können Sie dem Buch von Lothar Sachs (2004/2018) entnehmen.

5.3 Der Einstichproben-t-Test

Student's Einstichproben-t-Test wurde implizit bereits im ersten Abschnitt behandelt und ergibt sich ganz einfach aus dem Konfidenzintervall. Dessen Formel löst man nach $t_{P,n-1}$ auf und erhält

$$\frac{|\mu - \bar{x}| \cdot \sqrt{n}}{s} \leq t_{P,n-1}$$

Die Prüfung der Nullhypothese $H_0(\mu=0)$ ging in Abschnitt 5.1 via Konfidenzintervall vonstatten, indem man überprüfte, ob der hypothetische Wert $\mu=0$ im Intervall liegt oder nicht - mit anderen Worten, ob die beiden $\leq$-Relationen erfüllt sind oder nicht. Hier macht man nun das Gleiche: Man setzt in die letzte Formel den Wert $\mu=0$ ein und prüft, ob die Beziehung

$$t = \frac{|\bar{x}| \cdot \sqrt{n}}{s} \leq t_{P,n-1}$$

erfüllt ist. Dies ist stets genau dann der Fall, wenn das Konfidenzintervall "0" einschließt: Also wird für $t \leq t_{P,n-1}$ die Nullhypothese akzeptiert, im Fall $t > t_{P,n-1}$ wird sie abgelehnt. Erstes wieder mit dem Risiko β einer irrtümlichen Annahme der Nullhypothese, Zweites mit dem vorgewählten Risiko α einer irrtümlichen Ablehnung der Nullhypothese.

Im Beispiel der kardiologischen Studie wird in Tabelle 5 der berechnete t-Wert mit t=2.98 angegeben. Für die Konfidenz P=0.95 und fg=n-1=8 erhält man die kritische Schwelle aus der t-Tabelle mit $t_{0.95,8}=2.306 < t=2.98$, so dass H_0 abgelehnt wird. Nichts Neues, das weiß man schon aus Abschnitt 5.1, und tatsächlich: Das Konfidenzintervall sagt noch viel mehr aus als der t-Test, denn es gibt "nebenbei" noch eine Intervallschätzung für die mittlere Wirkung μ. Aus diesem Grund sollte man Konfidenzintervalle – sofern verfügbar - stets den korrespondierenden Testvarianten vorziehen oder optimal beide Methoden verwenden.

Dem Ergebnis t=2.98 ist in Abbildung 17a der Text "(p=0.017547)" angehängt. Dieser Wert ("p-Wert") kann als diejenige Irrtumswahrscheinlichkeit α aufgefasst werden, an der man vermöge des errechneten Wertes t=2.98 eben noch die Nullhypothese ablehnen könnte: Ist $p \leq \alpha=0.05$, so entspricht dies einer Ablehnung der Nullhypothese an der *a-priori*-Signifikanzschwelle $\alpha=0.05$. Umgekehrt ausgedrückt gibt der p-Wert die Wahrscheinlichkeit dafür an, dass das vorgefundene oder ein womöglich noch extremeres Stichprobenergebnis bei Gültigkeit der Nullhypothese H_0 "per Zufall" zustande kommen kann. Auf diese Weise können grundsätzlich alle Ergebnisse statistischer Testverfahren interpretiert werden.

Bei Testverfahren unterscheidet man einseitige und zweiseitige Tests. Die Konstruktion hier und in Abschnitt 5.1 arbeitet zweiseitig (zwei Grenzen des Intervalls), denn man möchte im Beispiel der Antihypertensivum-Studie Unterschiede in beiden Richtungen, also Blutdrucksenkungen und auch Druckerhöhungen erkennen. In diesem Skript wird durchgängig nur von zweiseitigen Tests gesprochen, die auch in fast allen Fällen indiziert sind. Eine der eher seltenen Ausnahmen von dieser Regel findet sich am Ende von Abschnitt 5.1 in Form eines einseitigen Binomial-Tests zu den Leukämie-Inzidenzen.

5.4 Der Zweistichproben-t-Test

Nach Lektüre des 2. Kapitels (Versuchsplanung) ist offensichtlich, dass der Kardiologe in Abschnitt 5.1 zusätzlich zu seiner eigentlichen Zielgruppe, die er mit der neuen Wirksubstanz behandelt, noch eine Kontrollgruppe mitführen sollte: Er könnte damit – bei vorausgesetzter Randomisierung - sicherstellen, dass in beiden Gruppen die gleichen äußeren Einflüsse wirksam sind und auf diesem Wege mögliche Unterschiede zwischen seiner neuen und zum Beispiel einer etablierten Standardtherapie als "echte" therapeutische Wirkungsunterschiede aufzeigen.

Der Einfachheit halber nehme man an, dass im Studienplan des Abschnittes 5.1 auch an eine Kontrollgruppe gedacht wurde. Diese Kontrollgruppe (ebenfalls mit n=9 Patienten, beachten Sie aber auch später den Abschnitt 5.10 zur Fallzahlberechnung!) mag zu folgenden Ergebnissen führen:

Systolischer Blutdruck in der Kontrollgruppe		
vor Behandlung	**nach Behandlung**	**Differenz**
180	170	10
180	165	15
175	160	15
175	170	5
190	185	5
185	180	5
180	170	10
175	170	5
180	180	0

Tabelle 6b: Ergebnisse der Hypertonie-Studie für die Kontrollgruppe

Als Auswertungsmethode kommt hier der sogenannte *Zwei-Stichproben-t-Test* in Frage:

Zweistichproben-t-Test

```
Neu (1):         n = 9   Xquer = 9.4   s = 9.5   s² = 90.3
Kontrolle (2):   n = 9   Xquer = 7.8   s = 5.1   s² = 25.7

Test auf Gauß-Verteilung:
Neu:       Kolmogoroff-Smirnoff's δ = 0.14 mit p > 0.2000: Ok
Kontrolle: Kolmogoroff-Smirnoff's δ = 0.26 mit p = 0.0711: ??

Test auf Homogenität der Varianzen (zweiseitig):
Prüfgröße F = 3.51 mit df = (8,8) und p = 0.094493: ??
>>  Empfohlen: Welch-Test für ungleiche Varianzen anwenden!

Zweiseitiger Test der Nullhypothese Ho(μ1=μ2):
Gleiche Varianzen:    t = 0.4643 mit df=16 und p = 0.64869
Ungleiche Varianzen:  t = 0.4643 mit df=12 und p = 0.65075

Konfidenzintervalle für die Differenz „Neu-Kontrolle“:
P = 0.95: [ -5.94 , 9.28 ]      P = 0.99: [ -8.82 ,12.15 ]
```

Abbildung 17b: Programmausgabe zur Hypertonie-Studie: Kontrolle vs. Novum

Abbildung 17b zeigt die vollständige Auswertung des Hypertonie-Versuchs mit zwei Parallelgruppen. Der erste Teil der Programmausgabe mit den Durchschnitten $\bar{x}$, den Standardabweichungen s, den Streuungen s^2 und den beiden Stichprobenumfängen n bedarf sicher keiner weiteren Erläuterung. Interessant und wichtig ist - vor dem eigentlichen Test! - eine Überprüfung der Voraussetzungen für den Zweistichproben-t-Test:

Das Programm überprüft für beide Stichproben die für parametrische Verfahren grundlegende Annahme, ob die Stichproben denkbarerweise aus Gauß-Verteilungen stammen oder nicht; dazu wird im Programm der Test von *Kolmogoroff und Smirnoff* durchgeführt. Die Nullhypothese lautet "Die Daten entstammen einer Gauß-Verteilung" und muss offenbar beibehalten werden: Die Verteilung der ersten Stichprobe ist unproblematisch, da der zugehörige p-Wert die kritische Schwelle von α=0.05 (besser: α=0.10, warum?) deutlich *nicht* erreicht, die zweite Stichprobe ist mit p=0.0711 gerade noch akzeptabel.

Als weitere Annahme für den Zweistichproben-t-Test ist die Annahme der *Homogenität der Varianzen* zu überprüfen: Dies erfolgt mit dem sogenannten F-Test, der im Beispiel gerade die Schwelle von α=0.10 unterschreitet, so dass man "sicherheitshalber", also konservativ keine Gleichheit der Varianzen unterstellen sollte. Der übliche Zweistichproben-t-Test sollte somit nur in der modifizierten Form des *Welch-Tests* angewendet werden:

Da man nach der letzten Überlegung von ungleichen Varianzen in den beiden Gruppen ausgehen muss, zieht man die entsprechende Prüfgröße

nach Welch zur Überprüfung der Nullhypothese $H_0(\mu_1=\mu_2)$ = H_0("Gleiche Behandlungswirkungen") heran. Der p-Wert beträgt p=0.650747 und überschreitet weit die kritische Grenze von α=0.05, so dass die Nullhypothese ganz eindeutig nicht abgelehnt werden kann. Zur Unterstützung dieser Entscheidung kann man auch die beiden Konfidenzintervalle für die Wirkungsdifferenz μ_1-μ_2 heranziehen, die bei einer Differenz $\overline{x}_1-\overline{x}_2=1.7$ ebenfalls eher auf eine Wirkungsgleichheit hinweisen.

Auf eine Angabe und Herleitung der zum Zweistichproben-t-Test verwendeten Formeln kann hier verzichtet werden, da die zum Verständnis parametrischer Methoden erforderlichen mathematischen Grundlagen bereits in den ersten Abschnitten dieses Kapitels behandelt wurden.

5.5 Der Wilcoxon-Mann-Whitney-Test

Der *Wilcoxon-Mann-Whitney-Test* (auch *U-Test, Rank-Sum-Test* oder eher unhistorisch *Mann-Whitney-Test* genannt: Der Test wurde zuerst 1945 von dem amerikanischen Chemiker Frank Wilcoxon (1892-1965) publiziert und 1947 von Mann und Whitney für ungleiche Fallzahlen verallgemeinert) ist ein Vertreter der Nicht-parametrischen Statistik und dient dem Vergleich von zwei Gruppen (zum Beispiel zweier Behandlungen "Novum" und "Standard") bei Vorliegen von wenigstens ordinalskalierten Merkmalen.

In einem Analgesie-Versuch soll ein Aktiv-Präparat mit einer Placebo-Behandlung verglichen werden, wozu einer Gruppe von n_1=10 Probanden Aktiv und einer zweiten Gruppe von n_2=9 Probanden Placebo verabreicht wird. Die Probanden sollen ihr Schmerzempfinden an Hand der Ordinalskala "0: Kein Schmerz, 1: Leichter Schmerz, 2: Mittlerer Schmerz und 3: Starker Schmerz" beurteilten. Im Versuch ergeben sich folgende Beurteilungen (auch "*Scores*" genannt):

Aktiv (n_1=10)	Placebo (n_2=9)
1	1
1	0
2	3
0	2
1	2
2	3
3	1
1	0
3	3
0	

Tabelle 8: Ergebnisse eines Analgesie-Versuches

Die Nullhypothese in diesem Beispiel lautet "Kein Wirkungsunterschied zwischen Aktiv und Placebo". Da die Zielgröße "Schmerzempfinden" keine quantitative, sondern eine ordinale Skala besitzt, scheidet ein parametrisches Testverfahren (zum Beispiel der t-Test) aus und es stellt sich die Frage nach einer Alternative:

Zur Erläuterung der Testkonstruktion wird - exemplarisch für die sogenannten Rangordnungsverfahren - etwas auf die mathematischen Hintergründe eingegangen, die sich in ähnlicher Form auch bei allen anderen Rangordnungstests wiederfinden. Leserinnen und Leser, die diese Dinge nur bedingt interessieren, können die entsprechenden Passagen gegebenenfalls auch übergehen. Unbedingt wichtig ist aber, sich über die Voraussetzungen des Tests, über die Nullhypothese und nicht zuletzt über die Interpretation der Testergebnisse im Klaren zu sein.

Zum rascheren Einstieg in die mathematische Problematik stelle man sich vor, dass nicht $n_1=10$ und $n_2=9$, sondern $n_1=n_2=3$ Probanden untersucht werden. Gemäß Anhang A.3 (Binomialkoeffizient) gibt es insgesamt $(n_1+n_2)!/(n_1!\cdot n_2!)=20$ Möglichkeiten, unterschiedliche Anordnungen von Aktiv (A) und Placebo (P) bezüglich der Schmerzäußerungen zu erhalten; zum Beispiel bedeutet die erste Anordnung in Tabelle 9, dass die drei kleinsten Werte von den Aktiv-Behandelten und die drei größten Werte aus der Placebo-Gruppe stammen. In Tabelle 9 sind alle kombinatorisch möglichen Anordnungen aufgelistet.

Nr	Anordnung	R_1	R_2	U_1	U_2	$\min(U_1,U_2)$
1	A A A P P P	6	15	9	0	0
2	A A P A P P	7	14	8	1	1
3	A A P P A P	8	13	7	2	2
4	A A P P P A	9	12	6	3	3
5	A P A A P P	8	13	7	2	2
6	A P A P A P	9	12	6	3	3
7	A P A P P A	10	11	5	4	4
8	A P P A A P	10	11	5	4	4
9	A P P A P A	11	10	4	5	4
10	A P P P A A	12	9	3	6	3
11	P A A A P P	9	12	6	3	3
12	P A A P A P	10	11	5	4	4
13	P A A P P A	11	10	4	5	4
14	P A P A A P	11	10	4	5	4
15	P A P A P A	12	9	3	6	3
16	P A P P A A	13	8	2	7	2
17	P P A A A P	12	9	3	6	3
18	P P A A P A	13	8	2	7	2
19	P P A P A A	14	7	1	8	1
20	P P P A A A	15	6	0	9	0

Tabelle 9: Schema zum Wilcoxon-Mann-Whitney-Test

Zur Konstruktion der Tabelle nummeriert man die 6 möglichen Rangplätze mit den Rangnummern von 1 bis 6. Für jede Anordnung addiert man die Rangnummern von A und erhält die Rangsumme R_1, entsprechend addiert man die Rangnummern für P und erhält die zweite Rangsumme R_2. (Zum Beispiel haben die Werte aus A in der ersten Anordnung die Nummern 1, 2 und 3, also Rangsumme $R_1=6$, die Werte der Placebo-Gruppe sind die drei größten mit den Rangnummern 4, 5 und 6, womit sich $R_2=15$ ergibt.)

Die nächsten Größen U_1 und U_2 dienen einer Standardisierung der Rangsummen R_1 und R_2. Dazu überlegt man sich, wie groß eine Rangsumme denn maximal werden kann; im Beispiel offenbar 15 (erste bzw. letzte Anordnung, maximale Trennung der beiden Gruppen), allgemeiner ist für n_1 und n_2 die maximale Rangsumme $R_{1(max)}$

$$R_{1(max)}=(n_2+1)+(n_2+2)+..+(n_2+n_1)=n_1 \cdot n_2+1+2+..+n_1=n_1 \cdot n_2+n_1 \cdot (n_1+1)/2$$

Der Ausdruck für die Summe 1+2+...+n=n(n+1)/2 geht angeblich auf den kleinen Gauß zurück, der in der Schule zur Übung die Zahlen von 1 bis 100 addieren sollte: An Stelle der langweiligen Rechnung überlegte er sich, dass ja 1+100=101, 2+99=101, ... und 50+51=101 ist, also die Rechnung $101 \cdot 50=(n+1) \cdot n/2=5050$ viel schneller zum richtigen Ergebnis führt.

Mit Hilfe von $R_{1(max)}$ kann man die Spalte $U_1=R_{1(max)}-R_1$ der Tabelle 9 definieren, ganz analog ergibt sich die Spalte $U_2=R_{2(max)}-R_2$. In der letzten Spalte wird der jeweils kleinere Wert von U_1 und U_2 in Form von $U=\min(U_1,U_2)$ aufgeführt.

In der letzten Spalte der Tabelle 9 treten offenbar *kleine* Werte von $U=\min(U_1,U_2)$ immer dann auf, wenn die A's und P's in den Anordnungen einigermaßen bzw. völlig getrennt sind. Sind die A's und P's "durchmischt", so erhält man in der letzten Spalte immer "große" Werte. Als Entscheidungsgrundlage bietet es sich also an, *kleine* Werte von U als Hinweis darauf aufzufassen, dass die Nullhypothese falsch ist, denn bei falscher Nullhypothese erwartet man eher eine ausgeprägte Trennung, mithin also keine Durchmischung der A's und P's. Zur formalen Definition der Irrtumswahrscheinlichkeit α zählt man nun einfach ab, wie oft welche Werte von $U=\min(U_1,U_2)$ in Tabelle 9 vorkommen:

$U=\min(U_1,U_2)$	n_U	h_U	H_U
0	2	2/20=0.10	0.10
1	2	2/20=0.10	0.20
2	4	4/20=0.20	0.40
3	6	6/20=0.30	0.70
4	6	6/20=0.30	1.00

Tabelle 10: Häufigkeiten der Prüfgröße U

Der extreme Wert von U=0, den man bei völliger Trennung der beiden Gruppen erhält, kommt in 2 von 20 Fällen vor, das sind 0.10=10%. Angenommen nun, die Nullhypothese H_0 ist richtig und Aktiv und Placebo wirken grundsätzlich völlig gleich, dann erhält man ein solches extremes Ergebnis immerhin "per Zufall" mit einer Wahrscheinlichkeit von 10%!

Vereinbart man ausnahmsweise eine Irrtumswahrscheinlichkeit α=0.10, so wird bei einer völligen Trennung der beiden Gruppen - also bei U=0 - die Nullhypothese mit der Irrtumswahrscheinlichkeit α=10% abgelehnt, denn ein extremes Ergebnis gibt ja eher einen Hinweis darauf, dass H_0 in Wirklichkeit falsch ist. Man spricht dann von einer unterschiedlichen Wirkung von Placebo und Aktiv und geht dabei das Risiko von 10% ein, dass diese Wirkung in Wahrheit gar nicht vorhanden ist, H_0 also trotzdem richtig ist und die vorgefundene extreme Trennung bzw. U=0 tatsächlich nur "per Zufall" zustande gekommen ist. (α=10% sollte ansonsten nur zum Beispiel bei einer Untersuchung von parametrischen Testvoraussetzungen verwendet werden oder um etwa Nebenwirkungen "auszuschließen", nicht aber bei Prüfung "auf Unterschiede". Bitte vergleichen Sie dazu auch den Abschnitt 5.8.)

Für größere Werte von n_1 und n_2 wird das Anlegen einer Tabelle für "kritische Werte" wie in Tabelle 10 kaum realisierbar sein, glücklicherweise gibt es dafür aber bereits in der Literatur ausführliche Tabellen (z.B. Lothar Sachs, Springer 2004/2018), die bei Bedarf verwendet werden können. Vom Rechnen mit Bleistift und Papier wird unbedingt abgeraten. Wenn man mit einem Computer-Programm arbeitet, sollte man dringend darauf achten, dass bei kleineren Fallzahlen nicht nur eine sogenannte approximative Testgröße, sondern, ähnlich wie gerade beschrieben, exakte Permutationstests verwendet werden: Erst ab ca. n=20 kann man eine angenäherte, auf die Standard-Gauß-Verteilung bezogene Testgröße Z verwenden (vgl. Abbildung 20). Letztere zeigt auch gleichzeitig die in der Statistik häufig geübte Praxis, eine Testgröße zu konstruieren, von der man zeigen kann, dass sie einer bereits bekannten Verteilung folgt (hier der Gauß-Verteilung!), falls die Nullhypothese richtig ist. Dieses Vorgehen findet man auch im nächsten Abschnitt wieder.

Da die Bleistift-und-Papier-Auswertung sehr aufwendig ist (gemeinsame Rangordnung, Rangnummern und -summen für beide Gruppen, minimales U feststellen), wurde zur Auswertung des Analgesie-Versuch ein Auszug der Ausgabe eines statistischen Programmpaketes – hier wieder **BiAS.** – verwendet; die Ergebnisse finden Sie in Abbildung 20.

Die p-Werte (vgl. dazu die Abschnitte 5.2-4) sind sehr groß und sprechen eindeutig für die Beibehaltung der Nullhypothese, denn erst für $p \leq \alpha$ würde man H_0 ablehnen (wie erwähnt, wählt man konventionellerweise α=0.05, 0.01 oder 0.001). Interessant sind noch die nicht-parametrischen Konfidenzintervalle ("Moses-Konfidenzintervalle"), die auf den Median der Differenzen zwischen den Gruppen bezogen sind. Diese Konfidenzintervalle

können ganz im Sinne der Abschnitte 4.3 und 5.1 (Schätzen und Testen!) verwendet und interpretiert werden.

Wilcoxon-Mann-Whitney-Test

```
Stichprobe 1: Rangsumme = 94.0
Stichprobe 2: Rangsumme = 96.0

Prüfgrößen:
U =   39.00  ( p = 0.660720, exakter Wert )
Z =   0.507  ( p = 0.612257, approximativ )

Moses-Konfidenzintervalle:
P = 0.9000:  (    -1.0000 ,     1.0000 )
P = 0.9500:  (    -2.0000 ,     1.0000 )
P = 0.9900:  (    -2.0000 ,     1.0000 )
```

Abbildung 20: Programmausgabe zur Auswertung des Analgesieversuchs

5.6 Der χ^2-Vierfeldertafel-Test

Der χ^2-Test (gelesen: χ^2=Chi-Quadrat) kann bei Vorliegen von Nominaldaten angewendet werden und dient vorwiegend dem Vergleich zweier Gruppen wie zum Beispiel zweier Therapien. Weitere Anwendungen des Tests ergeben sich bei Zusammenhangsanalysen von dichotomen Daten, womit der Test auch auf die *Ein*stichprobensituation anwendbar ist.

Einem Nephrologen fiel in seinen Patientenunterlagen eine Eigentümlichkeit auf, denn es schien, als seien Frauen mit Rhesusfaktor Rh+ im Vergleich zu den männlichen Patienten deutlich unterrepräsentiert. Er führt deshalb eine Untersuchung an 40 Frauen und 40 Männern durch und stellt zur ersten Auswertung die Ergebnisse in einer Vierfeldertafel zusammen:

	Rh+	**Rh-**	Summe
Frauen	**26**	**14**	40
Männer	**32**	**8**	40
Summe	58	22	80

Tabelle 11: Vergleich von Frauen und Männern bezüglich des Rhesusfaktors

Zur teststatistischen Analyse seiner Daten formuliert der Nephrologe die Nullhypothese $H_0(\theta_{weiblich}=\theta_{männlich})$: "Es besteht kein Unterschied zwischen Frauen und Männern bezüglich des Anteils Rhesusfaktor positiv". (Oder: "Es besteht kein Zusammenhang zwischen Rhesusfaktor und Geschlecht"!) Er erwartet natürlich, dass er die H_0 ablehnen kann, denn immerhin weisen in Tabelle 11 32/40=80% der Männer, aber nur 26/40=65% der Frauen Rhesusfaktor positiv auf. Als Test kommt für ihn nur der χ^2-*Vierfeldertafel-Test* in Frage.

Vor einer Auswertung der Daten in Tabelle 11 soll das Zustandekommen dieses Tests - stellvertretend für Tests auf Nominalskalenniveau - etwas näher beleuchtet werden. Auch diese Passage kann von mathematisch nicht begeisterungsfähigen Leserinnen und Lesern ohne Weiteres überschlagen werden; für diese ist aber sicher wieder die Interpretation der Ergebnisse am Ende des Abschnittes interessant.

	Rh+	**Rh-**	Summe
Frauen	$\mathbf{n_{11}}$ $\mathbf{\varphi_{11}}$	$\mathbf{n_{12}}$ $\mathbf{\varphi_{12}}$	$n_{1\bullet}= n_{11}+n_{12}$
Männer	$\mathbf{n_{21}}$ $\mathbf{\varphi_{21}}$	$\mathbf{n_{22}}$ $\mathbf{\varphi_{22}}$	$n_{2\bullet}= n_{21}+n_{22}$
Summe	$n_{\bullet 1}= n_{11}+n_{21}$	$n_{\bullet 2}= n_{12}+n_{22}$	$n = n_{\bullet\bullet} = \Sigma\, n_{ij}$

Tabelle 12: Symbolik zum χ^2-Vierfeldertafel-Test

In Tabelle 12 bedeutet z.B. n_{11} die Anzahl Frauen mit Rh+ und $n_{\bullet 1}= n_{11}+n_{21}$ die Gesamtanzahl Rh+. Falls die Nullhypothese "Kein Unterschied zwischen Männern und Frauen bezüglich des Rhesusfaktors" richtig ist, dann müssten die Verhältnisse $n_{11}/n_{1\bullet}$ (beobachteter Anteil Rh+ bei Frauen), $n_{21}/n_{2\bullet}$ (beobachteter Anteil Rh+ bei Männern) und $n_{\bullet 1}/n$ (Anteil Rh+ gesamt) im Wesentlichen gleich sein. Diese Überlegung kann man jetzt benutzen, um den untersuchten Zusammenhang präziser zu formulieren:

Falls H_0 richtig ist, so erwartet man im Feld z.B. links oben eine gewisse Anzahl φ_{11} (die sogenannte "*erwartete Häufigkeit*"), die sich aus der unter der H_0 erwarteten Beziehung $\varphi_{11}/n_{1\bullet}=n_{\bullet 1}/n=(n_{11}+n_{21})/n$, oder, aufgelöst nach $\varphi_{11}=n_{1\bullet}n_{\bullet 1}/n$, aus $\varphi_{11}/n_{1\bullet}=n_{\bullet 1}/n$ ergibt. Diese Anzahl φ_{11} *erwartet* man bei *richtiger* Nullhypothese, in der Untersuchung ergab sich aber die Anzahl n_{11} (die "beobachtete Häufigkeit").

Betrachtet man nun die Differenz $\delta=n_{11}-\varphi_{11}$, so spricht ein Wert nahe Null für die Nullhypothese (man erhält im Wesentlichen das, was man unter H_0 erwartet), ein dem Betrag nach großer Wert von δ spricht klar gegen die Nullhypothese, denn wenn man etwas ganz anderes beobachtet (nämlich n_{11}), als das, was man unter H_0 erwartet (das ist φ_{11}), so kann doch sicher nur die Annahme der Nullhypothese falsch sein. Eine analoge Argumentation gilt natürlich für die übrigen drei Felder der Tafel.

Damit ist die Grundlage für einen statistischen Test bereits vorhanden, denn wesentlicher Bestandteil der erforderlichen Prüfgröße sind die vier Differenzen n_{ij}-φ_{ij} für i=1,2 und j=1,2 in den vier Feldern. Da sich bei einer Summation die Differenzen gegenseitig aufheben, verwendet man im weiteren die jeweils *quadrierten* Differenzen $(n_{ij}-\varphi_{ij})^2$: Diese erkennt man in der nachfolgenden Formel wieder, die außerdem eine Standardisierung beinhaltet: Die Division der quadrierten Differenzen durch φ_{ij} ist plausibel, denn man muss den Unterschied der beiden Häufigkeiten n_{ij} und φ_{ij} an der Größenordnung φ_{ij} der beteiligten Werte relativieren. Die vier nunmehr standardisierten Abweichungen sind in allen vier Feldern von Interesse, so dass man endgültig die gewünschte Prüfgröße - in Vorgriff auf die mathematische Begründung mit χ^2 bezeichnet - als Summe der standardisierten Abweichungsquadrate aller vier Felder definieren kann:

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \varphi_{ij})^2}{\varphi_{ij}}$$

Kleine Werte dieser Prüfgröße χ^2 sprechen somit *für* die Nullhypothese H_0, *große* Werte von χ^2 eher *dagegen*, womit im letzteren Fall Anlass zu einer Ablehnung der Nullhypothese gegeben ist.

Wie ermittelt man nun den "kritischen Wert" für χ^2, dessen Überschreitung zur Ablehnung der Nullhypothese an der Signifikanzschwelle α führt?

Ein heuristischer, der Konstruktion des Wilcoxon-Tests analoger Ansatz führt zu einem sehr konservativen Test: Ermittle alle bei den gegebenen Randhäufigkeiten möglichen Tafeln (n_{11}=0,1,…,min($n_{1\bullet}$,$n_{\bullet 1}$)) und berechne dazu jeweils die Werte der Prüfgröße χ^2. Falls im konkreten Fall - etwa für Tabelle 11 - der berechnete χ^2-Wert zu den $\alpha\cdot 100\%$ größten gehört, so kann dies unter H_0 nur mit der Wahrscheinlichkeit α der Fall sein, womit - große χ^2-Werte sprechen gegen die H_0! - die Nullhypothese mit der Irrtumswahrscheinlichkeit α abzulehnen ist. Der englische Statistiker Egon S. Pearson (1895-1980) fand 1947 eine wesentlich effizientere mathematische Lösung:

Pearson fasste die Fragestellung als ein *Binomial-Problem* auf (vgl. Abschnitt 0.5), womit die Häufigkeiten $h_1=n_{11}/n_{1\bullet}$ und $h_2=n_{21}/n_{2\bullet}$ Schätzungen für die Parameter θ_1 und θ_2 (im Beispiel der Tabelle 11: θ_{weiblich} und $\theta_{\text{männlich}}$) darstellen. Unter der Nullhypothese ist $\theta=\theta_1=\theta_2$, die beste Schätzung für θ ist damit $h=(n_{11}+n_{21})/n=n_{\bullet 1}/n$ und, gemäß Abschnitt 0.5, ergibt sich daraus die empirische Varianz der beiden Schätzungen mit $h(1-h)/n_1$ bzw. $h(1-h)/n_2$: Dividiert man nun die Differenz $\delta=h_1-h_2$ durch ihre *bei Gültigkeit* der Nullhypothese gegebene Standardabweichung s_δ (letztere als Wurzel aus der Varianz der Differenz $\delta=h_1-h_2$: die Varianz einer Summe bzw. einer Differenz ist gleich der Summe der Varianzen!), so erhält man die standardisierte, angenähert Gauß-verteilte Prüfgröße

$$z = \delta / s_\delta = (h_1 - h_2) / \sqrt{h\cdot(1-h)\cdot(1/n_1 + 1/n_2)}$$

die man mit Hilfe der Gauß-Verteilung (Tabelle 4) beurteilen kann: Überschreitet der Betrag von z die Schwelle $u_P=u_{1-\alpha}$, so wird H_0 an der Signifikanzschwelle α abgelehnt. Mit einigem algebraischen Aufwand kann man zeigen, dass z^2 mit der Prüfgröße χ^2 identisch ist, womit auch der gesuchte Schwellenwert von χ^2_P mit $(u_P)^2$ identisch ist (vergleiche dazu die Tabellen 4 und 13!), und man erhält damit das mathematisch korrekte Procedere zur H_0-Prüfung.

Es lässt sich zeigen, dass die Prüfgröße χ^2 angenähert einer bekannten Wahrscheinlichkeitsverteilung folgt - der χ^2-Verteilung - falls die Nullhypothese *wahr* ist. Diese Näherung ist akzeptabel, falls $n\geq 20$ und $\varphi_{ij}\geq 5$ ist.

Im Falle des Vierfelder-Tests muss man nur einen speziellen Fall der χ^2-Verteilung kennen, die, ähnlich wie die t-Verteilung, neben P auch einen Freiheitsgrad df aufweist; für die vorliegende Vierfeldertafel genügen die Werte für df=1. Aus dieser χ^2-Verteilung mit df=1 ermittelt man per Integration - oder einfacher aus vorhandenen Tabellen, vgl. Tabelle 13 - den gewünschten "kritischen" Schwellenwert χ^2_P, um über eine Akzeptanz oder Ablehnung der Nullhypothese zu entscheiden.

Die Vierfelder-Tafel besitzt nur *einen* Freiheitsgrad: Sieht man die Randsummen einer Vierfelder-Tafel als gegeben an, so gibt es tatsächlich nur *eine* Möglichkeit, eine der vier Besetzungszahlen frei zu wählen, und nach freier Wahl (im Sinne einer zufällig gewonnenen Größe) kann/muss man wegen der gegebenen Randzahlen zwangsläufig die restlichen drei Häufigkeiten ausrechnen. Wegen der Bedingung vorgegebener Randhäufigkeiten wird der χ^2-Test auch als *bedingter Test* bezeichnet.

α = 1 - P	0.100	0.050	0.025	0.010	0.005	0.001
χ^2 mit df=1	2.706	3.841	5.024	6.635	7.879	10.828

Tabelle 13: Ausgewählte Werte der χ^2-Verteilung mit df=1

Im Beispiel des Rhesus-Faktors errechnet ein PC-Programm die Prüfgröße χ^2=2.257. Nach Vorgabe einer Irrtumswahrscheinlichkeit von α=0.05 erhält man aus Tabelle 13 die kritische Schwelle von $\chi^2_{0.95,1}$=3.841: Der errechnete Wert ist *kleiner* als diese Schwelle, so dass die Nullhypothese am Signifikanzniveau α=0.05 *nicht* abgelehnt werden kann. (χ^2=2.257 entspricht einem p-Wert von p=0.1330>0.05, der natürlich zum gleichen Ergebnis führt.) Der Unterschied in den prozentualen Anteilen von immerhin 15% kann also durchaus "zufällig" zustande gekommen sein, obwohl Frauen und Männern den gleichen Anteil Rh- bzw. Rh+ aufweisen. Wie bei allen statistischen Tests muss man bei einer Annahme der H_0 aber auch an den Fehler 2. Art der irrtümlichen Annahme denken, denn möglicherweise ist der vielleicht doch vorhandene Unterschied so klein, dass er mit n=80 Probanden (noch) nicht als "statistisch signifikant" aufzuzeigen ist!

In manchen Lehrbüchern findet man die Prüfgröße χ^2 auch in einer anderen Schreibweise mit der sogenannten *Yates-Korrektur*. Die Unterschiede in den Ergebnissen sind nur für kleine Fallzahlen relevant, ansonsten eher unwesentlich. Falls man ein PC-Programm zur Hand hat, sollte man ohnehin auf eine andere Testvariante, den sogenannten *Exakten Fisher-Test* ausweichen, der nicht nur - wie der χ^2-Test - für größere Fallzahlen berechnet werden darf, sondern auch für kleine Stichprobenumfänge exakte Ergebnisse liefert. Der Fisher-Test ist insbesondere bei Verletzung der Voraussetzung "n$\geq$20 und $\varphi_{ij}\geq$5" indiziert.

In Abschnitt 5.10 finden sich einige Hinweise auf die verallgemeinerte Form der Vierfelder-Tafeln: Dies sind sogenannte *Kontingenztafeln*, die eine Erweiterung des Tests auf mehr als zwei Zeilen und/oder Spalten darstellen. Berechnungsmöglichkeiten für alle Tests finden sich in **BiAS**..

5.7 Regressions- und Korrelationsrechnung

Das Gebiet der *Regressions- und Korrelationsrechnung* stellt ein wichtiges Methodenspektrum bereit, das in allen Anwendungsbereichen der Statistik eine zentrale Rolle einnimmt. Speziell im Umfeld der Medizin ist man mit diesen Methoden häufig konfrontiert, so dass es sich lohnt, auch hier einen Einblick nicht nur in die praktische Anwendung, sondern auch in die mathematische Struktur der Methoden zu gewinnen. Leserinnen und Leser mit eher peripherem Interesse an der Mathematik können die kurze Herleitung der Formeln wieder ignorieren, um sich stattdessen vielleicht eingehender mit den Anwendungen und den Beispielen zu beschäftigen.

In wissenschaftlichen Fachzeitschriften und in einschlägigen Lehrbüchern findet man häufig sogenannte *Regressionsgeraden* und *Korrelationskoeffizienten*, um die Abhängigkeit respektive den Zusammenhang zweier Größen beschreiben. In Abbildung 11 (Abschnitt 3.7) wurde bereits ein Beispiel zum Vergleich zweier Analysegeräte angesprochen, wozu vielfach die genannten Methoden herangezogen werden - auch wenn diese, wie später Abschnitt 6.2 zeigen wird, im Beispiel nicht ganz adäquat sind und zur Beurteilung nicht ausreichen. Vorweg sei betont, dass alle hier besprochenen Verfahren ausschließlich für quantitative Variablen definiert sind; auf Alternativen wird hingewiesen.

Die Regressions- und die Korrelationsrechnung stellen grundsätzlich unterschiedliche Methoden für inhaltlich unterschiedliche Fragestellungen bereit. In der Regressionsrechnung untersucht man die Abhängigkeit einer sogenannten *Zielgröße* Y von einer sogenannten *Einflussgröße* X. Dabei ist ganz eindeutig klar, welche der Größen von welcher abhängt, wie das Beispiel einer Dosis-Wirkungs-Beziehung zeigt: Die Serumkonzentration eines Pharmakons hängt eindeutig von der verabreichten Dosis ab, das Umgekehrte ist nicht denkbar.

In der Korrelationsrechnung existieren die beiden Begriffe Einflussgröße und Zielgröße nicht, denn dort spricht man nicht von einer *Abhängigkeit*, sondern von einem *Zusammenhang* zwischen zwei gleichberechtigten Merkmalen. Beispiele hierzu sind als Standard-Lehrbuchbeispiel der Zusammenhang zwischen Körperlänge und Körpergewicht oder, aus der Labormedizin, der Zusammenhang zwischen den beiden Schilddrüsenhormonen T3 (Trijodthyronin) und T4 (Thyroxin): In beiden Beispielen ist der Gedanke an eine eindeutige Abhängigkeit einer Variablen von der anderen abwegig. Zur formalen Unterscheidung verwendet man deshalb in der Korrelationsrechnung üblicherweise die beiden - gleichberechtigten - Symbole X_1 und X_2.

Zunächst wird ein näheres Augenmerk auf die Regressionsrechnung gerichtet. Das nächste Beispiel, in modifizierter Form dem Buch von Lorenz (1992) entnommen, ist typisch für eine - bereits angesprochene – Dosis-Wirkungs-Beziehung:

In einer Studie zur Untersuchung von Schadstoffbelastungen wird die Abhängigkeit der Bleikonzentration pro Kilogramm Nierenrinde (in mgPb) von der Menge Bleiacetat in der Nahrung untersucht (ebenfalls in mgPb angegeben). Wegen der gegebenen eindeutigen Abhängigkeitsrichtung kann in diesem Beispiel nur die Regressionsrechnung indiziert sein. Die graphische Darstellung der Ergebnisse wird mit einem *Scattergram* ("Punktwolke") vorgenommen, in das bereits die noch zu definierende *Regressionsgerade* eingetragen ist.

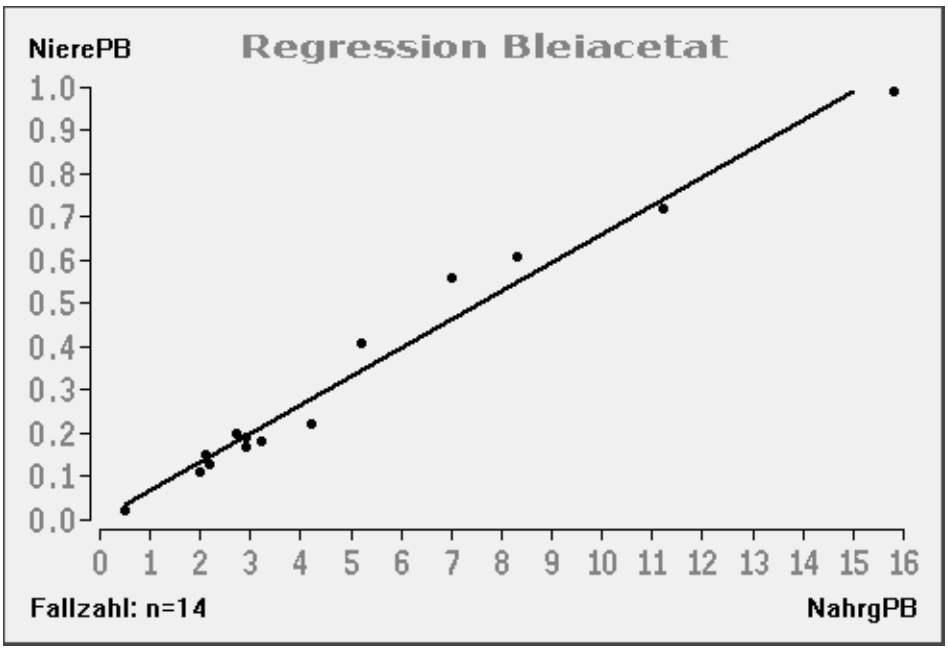


Abbildung 21: Abhängigkeit Nierenrinden- von Nahrungsbleigehalt

Als erste Aufgabe stellt sich die empirische Bestimmung der abgebildeten Regressionsgeraden: Die unten dargestellte *Methode der kleinsten Quadrate* (auch: *Ausgleichsrechnung,* engl. *Least Squares Method* oder *Fitting*) wird Carl Friedrich Gauß (1777-1855) zugeschrieben. Einige abkürzende Schreibweisen erleichtern die Ableitung der gesuchten Formeln:

Der Nahrungsgehalt wird mit X (Abszisse) und der Nierenrindengehalt des Bleiacetats mit Y (Ordinate) bezeichnet, die Stichprobenwerte sind jeweils Paare (x_i,y_i), wobei i wieder die Werte von 1 bis n=Stichprobenumfang annimmt. Der methodische Ansatz geht von den senkrechten Abständen der Punkte (x_i,y_i) von der - jetzt noch unbekannten - Regressionsgeraden aus; diese kann allgemein mit der Formel $Y=c+b\cdot X$ beschrieben werden. Ein solcher Abstand ist $(y_i-Y)=(y_i-(c+b\cdot x_i))$, wobei $c+b\cdot x_i$ der Wert der Geraden an der Stelle x_i ist. y_i ist der Y-Wert des Wertepaares (x_i,y_i). Natürlich müssen *alle* Wertepaare in die Berechnung eingehen, und so bildet man die Summe S der quadrierten (eine Begründung dafür folgt!) Abstände der Punkte von der Geraden:

$$S = \sum_{i=1}^{n} \{y_i - (c + b \cdot x_i)\}^2$$

Diese Summe sollte "möglichst klein" sein, denn die Gerade sollte die Punktwolke der n Wertepaare "möglichst gut" beschreiben in dem Sinne, dass die eben definierte Summe S der Abstandsquadrate so klein ist wie möglich.

Anschaulich kann man sich diese Minimierung so vorstellen, dass man sich Werte für die Steigung b und für den Achsenabschnitt c ausdenkt und diese so lange variiert, bis man glaubt das Minimum von S gefunden hat. Dieser Zugang ist - unabhängig vom Arbeitsaufwand - praktisch natürlich nicht umsetzbar, denn woran sollte man erkennen, ob man wirklich das gewünschte Minimum gefunden hat, oder ob es womöglich für ein weiteres Wertepaar c und b eine noch kleinere Summe S gibt?

Im Anhang wird zusammengefasst, wie man das Minimum einer quadratischen Funktion ermitteln kann (Abschnitt A.5): Man bildet die erste Ableitung, setzt diese gleich Null und löst die erhaltene Gleichung nach x auf. Die Summe S wurde bereits als quadratische Funktion (quadrierte Abstände!) definiert, denn ohne die Quadrierung der Differenzen wäre eine Minimalbetrachtung nicht möglich; von einer nur linearen Funktion - also ohne Quadrierung der Abstände - kann man bekanntlich kein Minimum finden. Im Anhang ist x die Größe, für die das Minimum von y=f(x) bestimmt werden soll, im vorliegenden Geradenproblem gibt es zwei Größen, die variiert werden können, um auf diesem Weg das Minimum von S=f(c,b) zu finden. Also bildet man zunächst die 1. Ableitung von S in Bezug auf c (dabei c wird als Variable aufgefasst) und ermittelt anschließend auch die 1. Ableitung von S bezüglich b (jetzt wird b als Variable aufgefasst). Man spricht dabei auch von sogenannten *partiellen Ableitungen*, da nicht nur eine Variable (im Anhang: x), sondern zwei Variablen (hier: c und b) eine Rolle spielen. Bei partiellen Ableitungen schreibt man nicht wie im Anhang z.B. dy/dx, sondern verwendet stattdessen ein "rundes d", das ist "∂".

Es ergeben sich die folgenden beiden Formeln, in denen die beiden ersten Ableitungen von S gleich Null gesetzt werden, denn auf diesem Weg erhält man auch im Anhang das Minimum:

$$\frac{\partial S}{\partial c} = 0 \text{ und } \frac{\partial S}{\partial b} = 0$$

Damit ist das gesteckte Ziel bereits erreicht, denn bekanntlich kann man zwei Gleichungen mit zwei Unbekannten (hier c und b!) auflösen und erhält damit eindeutige Lösungen für beide. Mit etwas Mühe und einigen algebraischen Umformungen erhält man den Achsenabschnitt c

$$c = \frac{\sum y_i}{n} - b \cdot \frac{\sum x_i}{n}$$

und die Steigung b der Regressionsgeraden

$$b = \frac{\sum (x_i \cdot y_i) - \frac{(\sum x_i) \cdot (\sum y_i)}{n}}{\sum x_i^2 - \frac{\sum x_i^2 - (\sum x_i)^2}{n}}$$

Die Steigung b wird üblicherweise auch als *empirischer Regressionskoeffizient* bezeichnet. Zur Berechnung empfiehlt sich die Verwendung eines statistischen Programmpaketes wie unter anderen das Programm **BiAS.**.

In der Regressionsrechnung kann man neben der nur deskriptiven Anwendung auch Nullhypothesen testen, denn man möchte in aller Regel untersuchen, ob die berechnete Steigung b der Regressionsgeraden nur "zufällig" von Null abweicht, in Wirklichkeit also vielleicht gar keine Abhängigkeit Y von X besteht. Sollte dies der Fall sein, so verläuft die Regressionsgerade in der Population parallel zur Abszisse, die Steigung β (griechischer Buchstabe β für den Populationsparameter, nicht zu verwechseln mit der Wahrscheinlichkeit β für den Fehler 2. Art!) ist also 0: Die Nullhypothese lautet deshalb $H_0(\beta=0)$, das Test-Procedere via Konfidenzintervall für die Steigung β bzw. t-Test für β erfolgt analog zu Abschnitt 5.1 und kann hier mit Hinweis auf einschlägige Programmpakete - zum Beispiel **BiAS.** - übergangen werden. Entspechendes gibt auch für den Test $H_0(\beta_0=0)$ für den Achsenabschnitt der Regressionsgeraden.

Im Beispiel der Bleibelastung der Niere errechnet man mit **BiAS.** c=0.0028, b=0.0658 und natürlich einen p-Wert, der sich mit $p<10^{-6}$ herausstellt. Die Nullhypothese wird abgelehnt, eine Abhängigkeit der Nierenrinden- von der Nahrungskonzentration von Bleiacetat ist damit "gesichert". (Bei $p<10^{-6}$ ist es fast müßig, über den möglichen Fehler 2. Art der irrtümlichen Ablehnung der H_0 zu diskutieren!) b=0.0658 kann im Beispiel gemäß der Definition einer Steigung (vgl. Anhang!) interpretiert werden: Erhöht man den Bleiacetatgehalt der Nahrung um eine Einheit, so steigt die Nierenrindenkonzentration im Mittel um den Betrag b=0.0658. Für den Achsenabschnitt c kann man ebenfalls eine Nullhypothese formulieren und testen, falls dies die Fragestellung erforderlich macht.

Bei Anwendung der Regressionsrechnung sollten keine auffälligen verteilungs- oder vielleicht auch messfehlerbedingte "Ausreißer" zu erkennen sein, außerdem sollten die Messwerte einigermaßen symmetrisch um die Gerade verteilt liegen. Zur Überprüfung dieser Voraussetzungen genügt in der Regel eine wenigstens graphische Beurteilung, formale Tests – dazu mehr in Abschnitt 5.2 - sind auch hier vorhanden. Eine Missachtung dieser Voraussetzungen ist immer ein "Kunstfehler"!

Nicht alle, prima vista an die lineare Regressionsrechnung erinnernden Fragestellungen sind auch für eine Auswertung damit geeignet. Im Beispiel der Dosis-Wirkungs-Beziehung etwa ist es bei pharmakologischen Untersuchungen oft von Interesse, die sog. *Mittlere Effektive Dosis ED50* zu berechnen: Dabei spielen aber in aller Regel nicht-lineare Beziehungen eine Rolle, so dass dazu alternative Methoden erforderlich sind. Ein einfaches, auch manuell durchführbares Verfahren dazu wird von Alexander et al. (J. Pharm. Toxicol. 1999, pp. 55-8) beschrieben, eine Berechnungsmöglichkeit dazu findet sich in dem Programm **BiAS.**.

Abbildung 22 beschäftigt sich mit dem Zusammenhang der beiden Schilddrüsenhormone T3 (Trijodthyronin, ng/ml) und T4 (Thyroxin, μg/dl). Da hier offenbar keine ausgesprochene Regressionssituation im Sinne einer eindeutigen Abhängigkeitsrichtung vorliegt, wurde zunächst die Größe T4 als Ziel- und T3 als Einflussgröße aufgefasst (durchgezogene Regressionsgerade) und die Analyse ein zweites Mal mit T4 als Einfluss- und T3 als Zielgröße durchgeführt (gestrichelte Gerade). Der Öffnungswinkel der beiden Regressionsgeraden ergibt ein Maß für die Enge des Zusammenhangs zwischen T3 und T4; dieses Maß wird im nächsten Absatz näher definiert. Interessant ist in Abbildung 22 das Hyperbelpaar: Die beiden Kurven geben Konfidenzgrenzen (hier wieder mit Konfidenz P=0.95) für die durchgezogene Regressionsgerade an; mit einer Sicherheit von P=0.95 verläuft die wahre, unbekannte Regressionsgerade der Grundgesamtheit in dem von den beiden Hyperbelästen begrenzten Raum. Die Graphik lässt erkennen, dass die Populations-Gerade somit auch durchaus parallel zur Abszisse verlaufen könnte, also keine Abhängigkeit der Größe T4 von T3 zu unterstellen wäre. Dieses Ergebnis erhält man auch mit Hilfe eines formalen Tests der Nullhypothese $H_0(\beta=0)$ (β ist die Steigung): Es ergibt sich als "p-Wert" die Überschreitungswahrscheinlichkeit p=0.078.

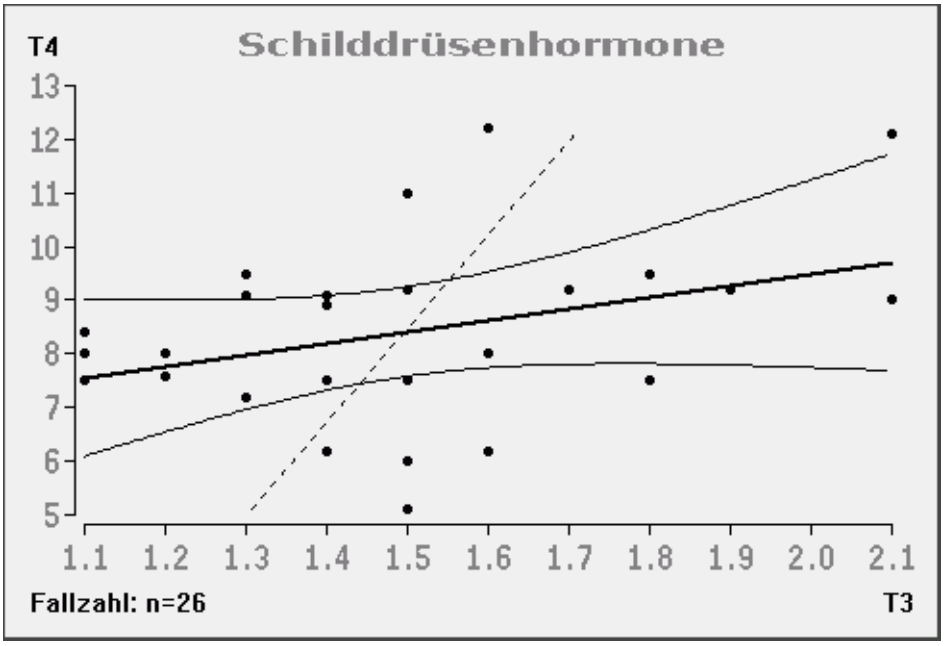


Abbildung 22: Korrelation der Schilddrüsenhormone T3 und T4

Mit dem Beispiel der Schilddrüsenhormone aus Abbildung 22 und mit dem bereits weiter oben dargestellten Vergleich zweier Analysemethoden (Abschnitt 3.7, Abbildung 11) liegen zwei typische Beispiele für die sogenannte *Korrelationsrechnung* vor, die auf den englischen Mathematiker Karl Pearson (1857-1936) zurückgeht. In der Korrelationsrechnung fasst

man - formal - einmal die eine, dann die andere der beiden Variablen als Einflussgröße auf, verwendet die jeweils andere als Zielgröße und berechnet mittels Regressionsrechnung die beiden jetzt sogenannten Steigungen b_{yx} und b_{xy} (gelesen: "b y auf x" bzw. "b x auf y"). Abbildung 22 zeigt die graphische Darstellung der beiden Regressionsgeraden. Proportional zum Öffnungswinkel dieser beiden Regressionsgeraden ist das geometrische Mittel $\sqrt{(b_{yx} \cdot b_{xy})}$ der beiden korrespondierenden Regressionskoeffizienten b_{yx} und b_{xy}, das, versehen mit dem gemeinsamen Vorzeichen von b_{yx} und b_{xy}, den sogenannten *Korrelationskoeffizienten r* definiert:

$$-1 \leq r \leq 1$$

Der Korrelationskoeffizient ist ein standardisiertes Maß für die Enge des Zusammenhangs zwischen den beiden betrachteten Variablen X_1 und X_2. Ist r>0, so spricht man von einem positiven Zusammenhang (je größer die eine Variable ist, umso größer wird auch die andere), ist r<0, so bedeutet dies einen negativen Zusammenhang (je größer die eine Variable ist, um so kleiner wird die andere). r=0 bedeutet, dass kein Zusammenhang besteht, in der Graphik stellt sich die Punktwolke "kugelförmig" dar. Die zugehörige Nullhypothese lautet $H_0(\rho=0)$ (griechischer Buchstabe ρ=rho für die Korrelation in der Grundgesamtheit). Auf die erforderlichen Formeln wird wieder mit Hinweis auf vorhandene Rechenprogramme - zum Beispiel **BiAS.** oder andere - verzichtet.

Im Beispiel des Vergleichs der beiden Analysemethoden aus Abschnitt 3.7 (Abbildung 11, "Scattergram") errechnet man nach Anwendung von **BiAS.** einen Korrelationskoeffizienten von r=0.9725 und einen p-Wert von $p<10^{-6}<<0.05$, womit die Nullhypothese an jeder der üblichen Signifikanzschwellen α abgelehnt werden kann: Es besteht also "zweifelsfrei" ein sehr enger Zusammenhang zwischen den Messwerten beider Geräte. Misst aber das eine Gerät das Gleiche wie das andere?

Dies ist entgegen vielfältiger Ansicht dadurch noch keineswegs gezeigt, wie im nächsten Kapitel ("Bland-Altman's Methodenvergleich") detaillierter dargestellt wird. Erste Hinweise auf eine Beurteilung gibt das nachfolgend definierte Bestimmtheitsmaß B, an dieser Stelle weiß man tatsächlich nur, dass ein - wie auch immer gearteter - Zusammenhang besteht, nicht aber, ob man in der Praxis das eine Messgerät, das vielleicht schneller, besser und kostengünstiger arbeitet, durch das andere ersetzen sollte.

Aus dem Korrelationskoeffizienten r kann man eine für die Interpretation der Ergebnisse sehr interessante Größe ableiten, die als *Bestimmtheitsmaß B* bezeichnet wird und definiert ist durch

$$B = \frac{SQ_{Gesamt} - SQ_{Rest}}{SQ_{Gesamt}} = \frac{\Sigma(y_i - \bar{y})^2 - \Sigma(y_i - (a + bx_i))^2}{\Sigma(y_i - \bar{y})^2} = r^2$$

SQ_{Gesamt} ist die Gesamtsumme der Quadrate in Y. Da im Allgemeinen nicht alle Punkte exakt auf einer Geraden liegen, ist SQ_{Rest} die "restliche" Summe der Abweichungsquadrate um die Regressionsgerade; B beschreibt somit den durch die Gerade erklärbaren Anteil der Y-Variabilität.

Liegen zum Beispiel alle Punkte auf einer Geraden, so ist $SQ_{Rest}=0$ und damit B=100%: Sämtliche Streuung in Y wird durch die Regression erklärt, zu dieser Erklärung sind keine äußeren Faktoren erforderlich. B kann nach Umformung einfacher über die Beziehung $B=r^2$ errechnet werden, woraus auch folgt, dass B für die beiden Geraden "Y auf X" und "X auf Y" den gleichen Wert besitzt. Im Beispiel des Analysenvergleichs ist $B=0.9725^2=0.9458$: 94.58% der Gesamtvariabilität des einen Gerätes kann durch die des anderen erklärt werden, mit dem Anteil von 5.62% gibt es einen äußeren Einfluss, der für die Abweichung der Geräte "verantwortlich" ist und dem dieser Varianzanteil zugeschrieben werden kann.

Die Multiple Regression untersucht die Abhängigkeit einer "Zielgröße" Y von nicht nur einer, sondern von mehreren "Einflussgrößen" X_i. Mit dem sogenannten "Abbau-Verfahren" werden alle Einflussgrößen, die nur marginalen bzw. keinen statistisch signifikanten Einfluss auf die Zielgröße besitzen, in der Reihenfolge ihrer Bedeutung aus dem Modell eliminiert, bis nur noch "wichtige", statistisch signifikante Variablen zur Erklärung der Zielgröße Y im Modell enthalten sind. Analog ist auch umgekehrt ein "Aufbau" des Modells möglich: Näheres dazu finden Sie in Abschnitt 6.10.

Die Multiple Korrelation untersucht analog den mehrdimensionalen Zusammenhang von mehreren Größen. Natürlich stehen auch dazu statistische Testverfahren und Konfidenzintervalle zur Verfügung.

In der Korrelationsrechnung muss man sich sehr vor *Scheinkorrelationen* in Acht nehmen. Das wohl bekannteste Beispiel ist die Scheinkorrelation zwischen der Geburtenhäufigkeit und der Anzahl vorhandener Klapperstörche, die man als statistisch hoch signifikant aufzeigen kann. Hier liegt sicher nicht der Fehler 1. Art vor (irrtümlich einen nicht vorhandenen Zusammenhang anzunehmen), vielmehr ist hier eine dritte Variable im Spiel, nämlich die Zeit bzw. die zunehmende Industrialisierung, die mit beiden Größen korreliert ist und den scheinbaren Zusammenhang zwischen Geburtenhäufigkeit und Klapperstorchanzahl produziert.

Als medizinisches Beispiel sei die Korrelation der Erythrozytenoberfläche und dem Hämoglobingehalt genannt, die verschwindet, wenn man die nominale Größe "Geschlecht" als dritte Variable berücksichtigt. Vor Scheinkorrelationen kann man selbst nach sorgfältigstem Nachdenken nicht sicher sein, denn immer kann es vielleicht noch unbekannte Faktoren geben, die möglicherweise die Scheinkorrelation bescheren: Grundsätzlich darf eine statistische Korrelation nur als ein formal-statistischer Zusammenhang verstanden werden, nicht aber als Dokumentation eines ursächlich-deterministischen Zusammenhanges. Zur Eliminierung des Einflusses von dritten Variablen kann die *Partielle Korrelation* verwendet werden.

Die parametrische Korrelationsrechnung besitzt die beiden nicht-parametrischen Analoga der *Spearman-* und der *Kendall-Korrelation*, womit auch ordinalskalierte Variablen auf ihre Korrelation untersucht werden.

Eine Korrelation auf Nominalskalenniveau ist mit Hilfe des χ^2-Vierfeldertests möglich. Die Nullhypothese lautet dabei "Es besteht kein Zusammenhang zwischen den beiden Nominal-Variablen", die Testgröße einschließlich des Testprocedere ist formal die gleiche wie bereits in Abschnitt 5.6 beschrieben wurde. Als Beispiel kann das gleiche wie im genannten Abschnitt dienen, wenn man dieses als Einstichprobenproblem interpretiert und damit die Korrelationssituation herstellt. Als Korrelationsmaß kann - analog zum Pearsonschen Korrelationskoeffizienten - bei dichotomen Daten "Pearson's φ-Koeffizient" verwendet werden.

Zur weitergehenden Lektüre sei wieder das mehrfach erwähnte Standardwerk von Lothar Sachs (Springer-Verlag 2004/2018) empfohlen, das zahlreiche Beispiele zu allen skizzierten Fragestellungen beinhaltet. In **BiAS.** finden sich Berechnungsmethoden zu allen genannten Verfahren.

5.8 Mehrfache Nullhypothesenprüfungen

Nur in seltenen Fällen steht in einer wissenschaftlichen Studie nur eine Nullhypothese zur Diskussion, sondern man muss in aller Regel mehrere statistische Hypothesen prüfen. Geht man bei der Auswertung seiner Daten also zwangsläufig mehrfach das Risiko ein, eine Nullhypothese irrtümlich abzulehnen (Fehler 1. Art, Irrtumswahrscheinlichkeit α!), so wird auch die Wahrscheinlichkeit, eine oder womöglich *mehrere* der getesteten Hypothesen irrtümlich abzulehnen, größer sein als bei *nur einer* Nullhypothese, jedenfalls also größer sein als die gewählte Irrtumswahrscheinlichkeit α für jeden einzelnen der N Tests. Damit tritt das Problem des *Multiplen Testens* zu Tage, das sich am besten am Würfelmodell veranschaulichen lässt:

Bekanntlich beträgt die Wahrscheinlichkeit p, mit einem Würfel eine "6" zu würfeln, p=1/6 - dies mag der irrtümlichen Ablehnung einer Nullhypothese entsprechen. Bei zweimaligem Würfeln (Testen) wird die Sachlage etwas komplizierter: Wie groß ist die Wahrscheinlichkeit, bei zweimaligem Würfeln eine Sechs zu erhalten, also im ersten *oder* im zweiten Wurf oder womöglich in *beiden* Würfen? (D.h.: Wie groß ist die Wahrscheinlichkeit, bei zwei Tests *mindestens einen* Test irrtümlich abzulehnen?) Die Wahrscheinlichkeit, in *beiden* Würfen *keine* Sechs zu würfeln, ist nach dem Multiplikationssatz (vgl. Abschnitt 0.2!) $p'=(5/6)^2=25/36$ (bei N Würfen entsprechend $p'=(5/6)^N$): Das Ereignis, bei zwei Versuchen *mindestens* eine "6" zu würfeln, besitzt also die komplementäre Wahrscheinlichkeit $p=1-p'=1-(5/6)^2=11/36$ (Entsprechendes gilt für N Versuche), die somit fast doppelt so hoch ist als bei nur einem Wurf. Oder intuitiv gesprochen: Man muss nur oft genug würfeln, irgendwann würfelt man eine "6" - und irgendwann wird man "mit Sicherheit" eine H_0 irrtümlich ablehnen!

Die Wahrscheinlichkeiten beim mehrfachen Testen von Nullhypothesen korrespondieren also mit den Wahrscheinlichkeiten beim mehrfachen Würfeln: Analog zum "Risiko" eine "6" zu würfeln geht man beim mehrfachen Nullhypothesentesten bei *jedem* einzelnen Test das Risiko α_{einzel} für den Fehler 1. Art ein. Und so ist beim Testen mehrerer Nullhypothesen die Gesamt-Wahrscheinlichkeit α_{gesamt}, irgendwelche der getesteten Nullhypothesen irrtümlich abzulehnen (Gesamt-Wahrscheinlichkeit für den Fehler 1. Art), ganz offensichtlich größer als nur α_{einzel} bei nur einem Test, und zwar umso größer, je mehr Nullhypothesen getestet werden: Die Formel ist die gleiche wie oben beim Würfeln, nämlich $\alpha_{gesamt}=1-P^N$, wobei $P=1-\alpha_{einzel}$ ist. Will man jetzt garantieren, dass bei N getesteten Nullhypothesen die Wahrscheinlichkeit α_{gesamt} für *mindestens eine* irrtümliche Ablehnung nicht "beliebig groß" wird, so setzt man in der Formel $\alpha_{gesamt}=1-P^N$ den gewünschten Wert von z.B. $\alpha_{gesamt}=0.05$ ein und löst die Beziehung nach P auf, denn dieses P bezieht sich gemäß Definition auf jeden *einzelnen* der N Tests: Damit kann man sich das P ausrechnen, das für jeden einzelnen der N Tests zu verwenden ist, um *insgesamt* das *fixierte* Gesamtrisiko $\alpha=\alpha_{gesamt}$ mit z.B. $\alpha=0.05$ zu gewährleisten. Diesem errechneten P, das für

jeden einzelnen der N Tests gilt, kommt deshalb – ebenfalls für jeden der N Einzeltests - ein $\alpha^*=\alpha_{einzel}=1-P$ zu, das man üblicherweise als *korrigierte* oder *adjustierte Irrtumswahrscheinlichkeit* α^* für jeden der N Individual-Tests bezeichnet. Der Wert von α, der für die gesamte Studie gelten soll, wird auch als *multiple Irrtumswahrscheinlichkeit* bezeichnet. Jeder Test, der formal bei $\alpha^*=\alpha_{einzel}$ abgelehnt wird, ist *global* bei α_{gesamt} signifikant.

$$\alpha = 1 - P^N = 1 - (1-\alpha^*)^N \Rightarrow \alpha^* = 1 - \sqrt[N]{1-\alpha} \approx \frac{\alpha}{N}$$

Die letzte Abschätzung der korrigierten Irrtumswahrscheinlichkeit $\alpha^*\approx\alpha/N$ geht auf Arbeiten des italienischen Mathematikers Carlo Emilio Bonferroni (1892-1960) zurück und wird deshalb auch *Bonferroni-Korrektur* genannt.

Angenommen, man muss in einer Untersuchung N=5 Nullhypothesen testen. Die Gesamt-Irrtumswahrscheinlichkeit soll α=0.05 betragen. Für jede der fünf Individual-Nullhypothesen bzw. für jeden der N Tests ist die korrigierte Irrtumswahrscheinlichkeit α^*=0.05/5=0.01 zu verwenden, um sicher zu sein, dass für die gesamte Studie eine multiple Irrtumswahrscheinlichkeit von α=0.05 eingehalten wird.

Dies alles gilt für Zielgrößen, deren Veränderung man gerne statistisch aufzeigen möchte und bei denen man sich immer gegen fälschlich "signifikante" Resultate absichern muss. Bei Nebenwirkungen dagegen darf man umgekehrt keine Bonferroni-Korrektur durchführen, um ja keine Nebenwirkungen zu übersehen: Im oben genannten Antihypertensivum-Beispiel könnte man etwa als Zielgröße die systolische Blutdrucksenkung verwenden, als Nebenwirkungsvariablen die Merkmale "Pulsfrequenz" und "diastolischer Blutdruck"; letztere sollten sich unter Medikation möglichst nicht verändern, die erste als Zielgröße natürlich schon. Da die α-Korrektur nur bei mehreren Zielgrößen relevant ist, entfällt hier eine Anwendung der Bonferroni-Korrektur, in Gegenteil verwendet man für Nebenwirkungsvariablen gelegentlich sogar eine Irrtumswahrscheinlichkeit von α=0.10, um ja keine Nebenwirkung zu übersehen (die Wahrscheinlichkeit β für den Fehler 2. Art wird dadurch "klein", vergleichen Sie dazu bitte Abschnitt 5.1!).

Bei Nebenwirkungsvariablen kann auch an eine Anwendung von *Äquivalenztests* gedacht werden. (Bei Äquivalenztests lautet die Nullhypothese nicht wie bei konventionellen Tests auf "Gleichheit", sondern umgekehrt auf "Ungleichheit": Eine solche am Signifikanzniveau α abgelehnte Nullhypothese lässt deshalb auf "Äquivalenz" schließen.) Äquivalenztests werden in diesem Skript jedoch nicht sehr ausführlich behandelt, einige Hinweise darauf und auf Tests auf "Nicht-Unterlegenheit" finden sich im nächsten Abschnitt 5.9.)

Neben der Bonferroni-Korrektur gibt es eine ganze Reihe weiterer interessanter Methoden (Holm-Prozedur, Hommel-Tests, Abschlusstests, hierarchische Prozeduren und viele andere mehr), die speziell bei "vielen" Zielgrößen (ab etwa 3-4) weniger konservativ sind. Darüber gibt die spezielle Literatur (zum Beispiel M. Horn und R. Vollandt, Multiples Testen, Fischer 1995) Aufschluss; das Prinzip ist bei allen Verfahren ähnlich.

5.9 Tests auf Äquivalenz und Nicht-Unterlegenheit

In den Abschnitten 5.4-6 wurden *Tests auf Unterschied* behandelt, denn das Ziel war, die Unterschiedlichkeit zum Beispiel zweier Therapieregime zu zeigen. Man spricht dabei auch von *Differenztests*, da man – zum Beispiel im Zweistichproben-t-Test – die Nullhypothese $H_0(\mu_1=\mu_2)$ betrachtet, die analog auch in der Form $H_0(\mu_1-\mu_2=0)$ geschrieben werden kann.

Tatsächlich ist es aber vielfach erwünscht, nicht den Unterschied, sondern die *Gleichwertigkeit* zweier Therapien oder Interventionen zu zeigen: In gewisser Weise ist dabei das Ziel, die Nullhypothese "zu beweisen". Nach Lektüre des Abschnittes 5.1 ist deutlich, dass eine nicht-abgelehnte Nullhypothese nicht auf die Gleichheit der Therapien schließen lässt, d.h. die Nullhypothese ist dadurch keineswegs "bewiesen": Ein nicht-signifikanter Unterschied ist nicht das Gleiche wie eine signifikante Übereinstimmung. Es stellt sich also die Frage nach einer geeigneten Methode.

Zur Illustration der Problematik stelle man sich im Rahmen der Antihypertensivum-Studie aus Abschnitt 5.4 vor, dass von dem langjährig bewährten Standard-Präparat nach Ablauf des Patentschutzes ein Generikum mit einem gegenüber dem Original um 60% reduzierten Apothekenpreis auf den Markt gebracht werden soll. (Ein Generikum ist eine exakte Kopie des Originalmedikamentes, es hat die gleiche galenische Form - Kapsel, Tablette, Lösung oder Zäpfchen - und natürlich denselben Wirkstoff wie das Originalprodukt.) Die Zulassungsbehörde BfArM verlangt nun aber einen *Nachweis auf Bioäquivalenz*, was bedeutet, dass der Hersteller eine Kontrollierte Klinische Studie mit dem Ziel durchführen muss, die therapeutische Äquivalenz der neuen Formulierung zu zeigen.

Die "klassische" Prüfung der Nullhypothese $H_0(\mu=0)$ im Einstichprobenfall per Konfidenzintervall wurde in Abschnitt 5.1 dargestellt. Analog kann man auch – vgl. dazu Abschnitt 5.4 – im Zweistichprobenfall zum Test auf Unterschied ein Konfidenzintervall für die Differenz $\mu_1-\mu_2$ berechnen (vgl. Abbildung 17b) und die Prüfung von $H_0(\mu_1-\mu_2=0)$ mit Hilfe dieses Konfidenzintervalls durchführen: Ist der Wert 0 im Intervall enthalten, so wird H_0 akzeptiert, ansonsten mit Irrtumswahrscheinlichkeit α abgelehnt.

Zur Durchführung eines *Tests auf Äquivalenz* formuliert man die Nullhypothese $H_0(|\mu_1-\mu_2|>\delta)$, wobei δ diejenige Grenze darstellt, für die man noch die Gleichwertigkeit der beiden Behandlungen akzeptieren kann: Mit anderen Worten ist δ ist *maximale irrelevante Differenz* zwischen den beiden Behandlungen. Die analoge Alternativhypothese H_A lautet damit $H_A(|\mu_1-\mu_2|\leq\delta)$, womit man gezeigt hätte, dass der Unterschied *kleiner* ist als die irrelevante Differenz δ und schließt in diesem Sinne auf Äquivalenz. Eine kritische Stimme zur Festlegung von δ findet sich bei Meyer (2008).

Das dazu erforderliche Testprocedere ist sehr anschaulich: Liegt das zweiseitige $(1-2\alpha)\cdot 100\%$-Konfidenzintervall (für $\alpha=0.05$ ist $P=(1-2\alpha)\cdot 100\%=90\%$!) für $\mu_1-\mu_2$ *gänzlich* im *Äquivalenzbereich* $[-\delta,+\delta]$, so wird die Hypothese $H_0(|\mu_1-\mu_2|>\delta)$ bei α abgelehnt, in *allen anderen Fällen* akzeptiert: Dieser sog. *Intervall-Inklusionstest* entspricht zwei *ein*seitigen t-Tests für $H_{01}(\mu_1-\mu_2>\delta)$ und $H_{02}(\mu_2-\mu_1>\delta)$ jeweils bei $\alpha(=0.05)$, die *beide* signifikant sein müssen und deshalb keine α-Korrektur nach 5.8 erforderlich machen.

Bei Prüfung auf *Nicht-Unterlegenheit* ("*Non-Inferiority*", *ein*seitiger Test!) betrachtet man die beiden Hypothesen $H_0(\mu_1-\mu_2>\delta)$ resp. $H_A(\mu_1-\mu_2\leq\delta)$ und verwendet ein nur *ein*seitiges $(1-\alpha)\cdot 100\%$-Konfidenzintervall für $\mu_1-\mu_2$.

In diesem Skriptum werden durchgängig nur klassische Tests auf Unterschied behandelt, jedoch kann man aus den zugehörigen Konfidenzintervallen mit Hilfe eines Äquivalenzbereiches stets einen Äquivalenztest ableiten. Theoretisches dazu findet sich reichlich bei Wellek (1994), Praktisches im Programm **BiAS.**.

5.10 Fallzahlberechnungen

In Abschnitt 4.5 wurde eine Formel zur Fallzahlberechnung für das Konfidenzintervall im "Ein-Stichproben-Fall" abgeleitet. Obwohl man mit einem Konfidenzintervall die Nullhypothese $H_0(\mu=0)$ prüfen kann, gilt die Fallzahl-Formel trotzdem nicht für die Testsituation, denn hierzu spielt auch die Wahrscheinlichkeit β für den Fehler 2. Art eine wesentliche Rolle; vgl. Abschnitt 5.1. Da sich aus dem Test-Procedere via Konfidenzintervall äquivalent der Einstichproben-t-Test ergibt (vgl. ebenfalls Abschnitt 5.1), gilt die nachfolgend abgeleitete Fallzahl-Formel sowohl für den Test via Konfidenzintervall als auch für den Einstichproben-t-Test:

Eine Formel für das Konfidenzintervall bei bekanntem σ wurde in Abschnitt 4.3 hergeleitet, diese lautet

$$\bar{x} - u_P \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + u_P \cdot \frac{\sigma}{\sqrt{n}}$$

Daraus ergibt sich nach Umformung der Ausdruck

$$|\bar{x} - \mu| \leq u_P \cdot \frac{\sigma}{\sqrt{n}} - u_P \cdot \sigma_{\bar{x}}$$

Dieser Sachverhalt ist bereits von der Gauß-Verteilung her bekannt (vgl. Abschnitt 4.2): $\bar{x}$ besitzt die Standardabweichung $\sigma_{\bar{x}}$, so dass der Anteil P aller Werte $\bar{x}$ zwischen $\mu\pm u_P\cdot\sigma_{\bar{x}}$ liegt. Also kann - an Stelle eines Tests via Konfidenzintervall - zum Test der Nullhypothese $H_0(\mu=0)$ in der letzten Formel $\mu=0$ gesetzt und H_0 mit der Irrtumswahrscheinlichkeit $\alpha=1-P$ abgelehnt werden, falls der Betrag von $\bar{x}$ die Grenze $K=u_P\cdot\sigma_{\bar{x}}$ überschreitet. Dies zeigt die linke Verteilung der nächsten Abbildung 23.

Äquivalent dazu kann man bei bekanntem σ aus der letzten Formel analog zum Einstichproben-t-Test den sogenannten *Gauß-Test* ableiten, der jedoch praktisch nicht von Bedeutung ist. Die Prüfgröße des Gauß-Tests ist $u=|\bar{x}|/\sigma_{\bar{x}}$ (unter der Nullhypothese ist $\mu=0$) und muss zur Ablehnung der Nullhypothese den Wert u_P überschreiten.

Falls das unbekannte μ in Wahrheit gleich $\delta>0$ ist (die Nullhypothese ist falsch, Analoges gilt für $\delta<0$), so ist offensichtlich ein Anteil β aller $\bar{x}$-Werte dem Betrag nach kleiner oder gleich $K=u_P\cdot\sigma_{\bar{x}}$. Dies bedeutet dann

aber nach der bekannten Entscheidungsregel eine irrtümliche Annahme der Nullhypothese; diese irrtümliche Annahme wird bekanntlich als "Fehler 2. Art" bezeichnet und besitzt die Wahrscheinlichkeit β. Die Situation wird in Abbildung 23 graphisch dargestellt.

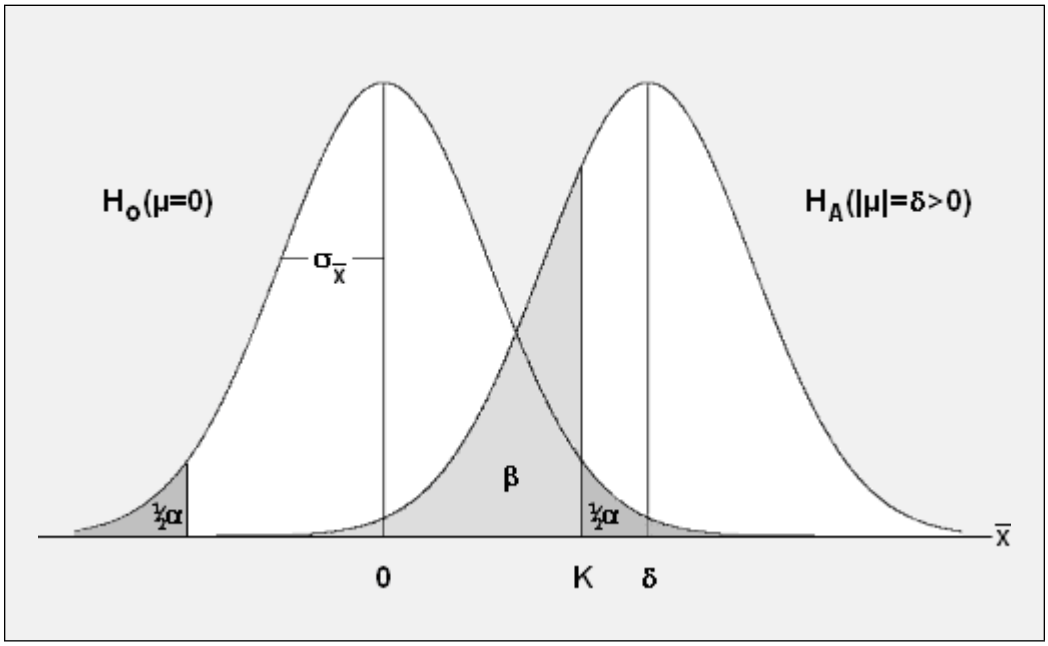


Abbildung 23: Skizze zur Fallzahlberechnung

Aus der Sicht der linken Verteilung (Nullhypothese!) errechnet sich K nach der Formel

$$K = u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

Aus Sicht der rechten Verteilung (Alternativhypothese!) errechnet sich K nach der Formel

$$K = \delta - u_{1-2\beta} \cdot \frac{\sigma}{\sqrt{n}}$$

Bitte beachten Sie, dass hier die beiden Schranken $u_{1-\alpha}$ und $u_{1-2\beta}$ *zwei*seitig definiert sind: Manche Lehrbücher weichen von dieser Notation ab und verwenden in den letzten beiden Formeln eine Schreibweise mit *ein*seitigen Schranken $u_{1-\alpha/2}$ und $u_{1-\beta}$. Aus Konsistenzgründen und zur Vereinfachung werden in diesem Skript konsequent nur *zweiseitige Schranken* verwendet.

Bei bekanntem σ besteht offenbar ein Zusammenhang zwischen den Größen α, β, δ und n: Gibt man drei davon vor, so kann man sich die jeweils vierte errechnen. Bei der Planung einer Studie wird man sich

natürlich nach der "richtigen" Fallzahl n fragen und kann zunächst α und β festlegen. Die dritte Größe δ ist der Betrag der *minimalen medizinisch relevanten Differenz* des wahren Mittelwertes μ von dem hypothetischen Wert 0, die man in der Studie als "statistisch signifikant" aufzeigen möchte: Diese Differenz δ kann nur unter medizinischen Gesichtspunkten definiert werden! Bei nunmehr fixierten Werten von α, β und δ erhält man durch Gleichsetzen der letzten beiden Formeln die Beziehung

$$u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} = \delta - u_{1-2\beta} \cdot \frac{\sigma}{\sqrt{n}}$$

Diese Formel löst man nach n auf und kann dadurch bei gegebenen α, β und δ die zum "Nachweis" der medizinisch relevanten Differenz δ erforderliche Fallzahl n errechnen. Zunächst ergibt sich nach Umformung

$$\delta = (u_{1-\alpha} + u_{1-2\beta}) \cdot \frac{\sigma}{\sqrt{n}}$$

und daraus wiederum

$$n = \left(\frac{(u_{1-\alpha} + u_{1-2\beta}) \cdot \sigma}{\delta} \right)^2$$

Bei unbekanntem σ, das beim Einstichproben-t-Test dem σ der Paardifferenzen entspricht, setzt man - etwa aus Pilotstudien oder aus der Literatur bekannt - als Schätzung für σ die empirische Standardabweichung s ein und erhält auf diesem Wege eine Näherung für n. Schätzt man mangels anderer Informationen die Spannweite R der zu erwartenden Daten, so kann man wie in Abschnitt 4.5 die grobe Abschätzung $s \approx R/4$ verwenden.

Im bisher diskutierten Einstichprobenfall muss man wie in Abschnitt 4.5 die gewünschte Irrtumswahrscheinlichkeit α und jetzt zusätzlich die gewünschte Wahrscheinlichkeit β für den Fehler 2. Art festlegen. Übliche Werte für β sind 0.10 und 0.20; in Abhängigkeit von der konkreten Fragestellung ist jedoch in Erwägung zu ziehen, für β den gleichen Wert zu wählen wie für die Irrtumswahrscheinlichkeit α. Die Standardabweichung s wird wie in Abschnitt 4.5 aus z.B. Vorstudien geschätzt, so dass jetzt nur noch die minimale, medizinisch relevante Differenz δ, die im Test "nachgewiesen" werden soll, zu definieren ist: Erachtet der Kardiologe im Beispiel des Abschnittes 5.1 etwa eine Blutdruckveränderung von 10 mmHg als medizinisch relevant, so setzt man zur Planung des Stichprobenumfangs in der eben abgeleiteten Fallzahl-Formel für δ den Wert δ=10 ein. Aufgrund des Pilotversuches kennt man eine Schätzung für σ (hier also s=9.5, vgl. Abbildung 17a), die Werte von u für zum Beispiel α=0.05 und β=0.10 sind aus Tabelle 4 abzulesen, es $u_{1-0.05}=u_{0.95}=1.96$ und $u_{1-2\cdot0.10}=u_{0.80}=1.28$. Die Fallzahl-Formel ergibt - aufgerundet - einen Stichprobenumfang von n=10, womit die kardiologische Studie durchaus eine angemessene Fallzahl aufweist.

Die Fallzahl-Formel gilt für zweiseitige Fragestellungen (vgl. Abschnitt 5.3 und 5.4): Die Nullhypothese lautet $H_0(\mu=0)$, die Alternativhypothese ist

$H_A(\mu \neq 0)$, womit also Veränderungen in beide Richtungen entdeckt werden können. (Die Länge L des Konfidenzintervalls in Abschnitt 4.5 entspricht $\delta=L/2$; dort wurde implizit $\beta=0.5$ bzw. $u_{1-2\beta}=u_{1-1}=u_0=0$ angenommen!).

Die letzte Formel gilt nur bei bekannter, nicht aber bei unbekannter Varianz σ^2 und unterschätzt speziell bei kleinen Fallzahlen systematisch die zum "Nachweis" der medizinisch relevanten Differenz δ tatsächlich erforderliche Fallzahl n. Präzisere Formeln für n, die jedoch ohne ein einschlägiges Computerprogramm kaum zu berechnen sind, wurden in einigen Programmpaketen (so auch in **BiAS**.) für den vorliegenden Fall und für viele weitere Fragestellungen implementiert. Die Ableitung der letzten Formel besitzt somit zwar nur exemplarischen Charakter, ist aber bei resultierenden Fallzahlen n ab n>20 recht genau.

Ganz ähnlich ergibt sich eine Fallzahl-Formel für den Zweistichproben-t-Test, die hier ohne Ableitung angegeben wird. Zur Übung kann man die Formel nach dem obigen Vorbild nachvollziehen, wenn man sich analog zum Einstichproben-t-Test einerseits an Stelle von μ und $\bar{x}$ mit $\delta=\mu_1-\mu_2$ und $d=\bar{x}_1-\bar{x}_2$ und andererseits bei unterstellt gleichen Standardabweichungen $\sigma=\sigma_1=\sigma_2$ mit der Standardabweichung σ_d der Differenz d beschäftigt:

$$n_1 = n_2 = 2 \cdot \left(\frac{(u_{1-\alpha} + u_{1-2\beta}) \cdot \sigma}{\delta} \right)^2$$

Weitere Ableitungen von Fallzahl-Formeln für die übrigen, in diesem Skriptum vorgestellten Tests sind mathematisch zum Teil recht aufwendig und werden hier nicht dargestellt, zumal diese zum Verständnis des statistischen Prinzips auch nichts Neues beitragen: Zur Vervollständigung der besprochenen Tests werden deshalb im Folgenden nur die relevanten Formeln zur Fallzahlberechnung angegeben. Wie erwähnt, stehen für die praktische Berechnung auch einschlägige Programme (zum Beispiel **BiAS.**) zur Verfügung. Zur Terminologie ist noch zu bemerken, dass $1-\beta$ auch als "Power" bzw. "Teststärke" eines statistischen Tests bezeichnet wird.

Zur Fallzahlberechnung für den Wilcoxon-Mann-Whitney-U-Test wird die zuletzt angegebene Formel häufig als "Näherung" herangezogen, obwohl diese nur für die parametrische Fragestellung der Zweistichproben-Situation zutrifft. Besser verwendet man den Mann-Whitney-Schätzer $p=P(X_1<X_2)$, der die geschätzte bzw. gewünschte Wahrscheinlichkeit dafür angibt, dass ein Wert aus Population 1 einen kleineren Wert aufweist als ein Wert aus Population 2; $p=P(X_1<X_2)$ kann günstigstenfalls aus Vorstudien (zum Beispiel wieder mit **BiAS.**) ermittelt werden. Die folgende Formel gilt für die zweiseitige Fragestellung:

$$n_1 = n_2 = \frac{(u_{1-\alpha} + u_{1-2\beta})^2}{6 \cdot (p - 0.5)^2}$$

Zur Fallzahlberechnung für den χ^2-Vierfeldertafel-Test gibt man die minimale, medizinisch relevante Differenz $\delta=|\theta_1-\theta_2|$ durch die beiden geschätzten oder wenigstens "vermuteten" Erfolgsraten θ_1 und θ_2 vor und berechnet mittels $\bar{\theta}=(\theta_1+\theta_2)/2$ die zur Ablehnung der Nullhypothese $H_0(\theta_1=\theta_2)$ erforderlichen Stichprobenumfänge n_1 und n_2; die zweiseitige Alternativhypothese lautet $H_A(\theta_1\neq\theta_2)$:

$$n_1 = n_2 = \left(\frac{u_{1-\alpha} \cdot \sqrt{2 \cdot \bar{\theta} \cdot (1-\bar{\theta})} + u_{1-2\beta} \cdot \sqrt{\theta_1 \cdot (1-\theta_1) + \theta_2 \cdot (1-\theta_2)}}{|\delta|} \right)^2 + \frac{2}{|\delta|}$$

Zum Binomialtest formuliert man die Nullhypothese $H_0(\theta=\theta_0)$ und möchte diese ablehnen, falls der wahre Parameter θ um mindestens die minimale, medizinisch relevante Differenz δ von dem hypothetischen Wert θ_0 abweicht. Daraus ergibt sich die einseitige (!) Fallzahlberechnung für den Binomial-Test:

$$n = \left(\frac{u_{1-2\alpha} + u_{1-2\beta}}{2 \cdot (\arcsin\sqrt{\theta_0} - \arcsin\sqrt{\theta_0 + \delta}} \right)^2$$

In den Abschnitten 0.5 und 5.1 wurde ein Beispiel zur Leukämie-Inzidenz im Umkreis eines Atomkraftwerkes diskutiert; die Leukämierate in der Bundesrepublik wurde mit θ_o=0.000104 angegeben. Ein Untersucher vermutet eine Verdopplung der Leukämierate in Kraftwerksnähe (auf θ_{AKW}=0.000208, also eine "nachzuweisende" relevante Differenz von $\delta=\theta_{AKW}-\theta_o$=0.000104) und möchte diese in einem *ein*seitigen Test mit der Alternativhypothese $H_A(\theta>\theta_o)$ auch als statistisch signifikant aufzeigen: Er entnimmt für z.B. α=0.05 und β=0.10 der Tabelle 4 die beiden Werte $u_{1-2\alpha}=u_{0.90}$=1.644854 und $u_{1-2\beta}=u_{0.80}$=1.281552 (bitte beachten Sie, dass in diesem Skriptum aus systematischen Gründen nur zweiseitige "kritische Größen" u verwendet werden!) und errechnet mit Hilfe der letzten Formel den erforderlichen Stichprobenumfang n mit n=119967(!). Die arcsin-Transformation als Umkehrfunktion des trigonometrischen Sinus steht auf fast allen Taschenrechnern als "$\sin^{-1}$" zur Verfügung.

Unabhängig vom letzten Beispiel sollte abschließend noch bemerkt werden, dass bei der Planung einer Studie mit mehreren, zum Beispiel N Nullhypothesenprüfungen die Fallzahlberechnungen nicht für das multiple Signifikanzniveau α, sondern für jeden einzelnen der N Tests für das Bonferroni-korrigierte Signifikanzniveau $\alpha^*=\alpha/N$ durchgeführt werden (zur Bonferroni-Korrektur vgl. Abschnitt 5.8). Verwendet man als Fallzahl für die gesamte Studie das Maximum der errechneten N individuellen Fallzahlen, so ist gewährleistet, dass die "Power" ("Teststärke") des multiplen Test mindestens gleich 1-β ist bzw. dass die Wahrscheinlichkeit für den Fehler 2. Art den Wert β nicht überschreitet.

5.11 Ausblick

Sowohl die parametrische als auch die nicht-parametrische Statistik bieten eine reiche Fülle von weiteren Verfahren an, die auf viele und viel allgemeinere Fragestellungen anwendbar sind als die bisher angesprochenen. Aus dem bisher geläufigen Stand der Darstellung sind einige Ergänzungen und Verallgemeinerungen erwähnenswert:

Der *Wilcoxon-matched-pairs-Test* (auch: *Wilcoxon-Signed-Rank-Test*) stellt eine nicht-parametrische Analogie zum Einstichproben-t-Test dar. Die Bezeichnung "matched-pairs" weist darauf hin, dass pro Merkmalsträger jeweils zwei Werte (zum Beispiel vor und nach Intervention o. ä.) gemessen werden und deren Paardifferenz in die Auswertung eingeht. Wie bereits erwähnt, sind nicht-parametrische Verfahren insbesondere dann von Interesse, wenn die grundlegende Voraussetzung "Gauß-Verteilung" nicht unterstellt werden kann und/oder keine quantitative Skala vorliegt.

Außer dem behandelten parametrischen und nicht-parametrischen Vergleich zweier Stichproben (Abschnitte 5.4 und 5.5, t-Test und U-Test) besteht auch die Möglichkeit, *mehrere* Stichproben in *einem* in sich abgeschlossenen sog. *Global-Test* zu vergleichen (das sind: *Varianzanalyse, Kruskal-Wallis-Test*) und im Anschluss daran paarweise Vergleiche mit speziellen Methoden durchzuführen. Damit entfällt eine α-Korrektur nach z.B. Bonferroni (Abschnitt 5.8), die besonders bei mehr als drei paarweisen Vergleichen zu recht konservativen Testentscheidungen führen kann.

Die Vierfeldertafel-Tests für Nominaldaten (χ^2- und Fisher-Test) sind ebenfalls auf beliebig viele Zeilen und Spalten verallgemeinerbar (dabei spricht man von *Kontingenztafeln*), wobei hierzu wiederum spezielle Testverfahren für Kontingenztafeln mit kleinen Besetzungszahlen existieren (das sind: *Haldane-Dawson-Test* oder auch *Exakte Tests*). Das Buch von Lothar Sachs (Springer 2004/2018) ist "für alle Fälle" auch hier eine unerschöpfliche Fundgrube.

Ein Ansatz zum Binomial- und Vorzeichentest wurde bereits beschrieben (vgl. Beispiel Leukämie-Inzidenz, Abschnitte 0.5 und 5.1). Mit Abschnitt 6.1 (Qualitätssicherung) führt ein Test auf Trend von Kontrollmessungen im Labor ebenfalls zum *Binomial-Test*.

Besonders vielfältig sind Methoden zur Regressions- und Korrelationsrechnung. Außer Geraden kann man auch *Polynomiale Modelle* untersuchen, man kann *Loglineare Modelle* und *Exponentialmodelle* betrachten, wenn der Zusammenhang - zum Beispiel altersabhängig - nicht mit der Annahme der Linearität vereinbar ist und kann auch die Abhängigkeit einer Zielgröße von mehreren Einflussgrößen analysieren.

Alle eben angesprochenen verallgemeinerten Methoden sind mit dem Programm **BiAS.** durchführbar. Im Handbuch finden sich zahlreiche konkrete Beispiele mit empirischen Daten und den Berechnungsresultaten.

Eine eingehende Lektüre des 5. Kapitels schafft nach Ansicht des Autors ausreichende Grundlagen für eine Auseinandersetzung auch mit "höheren" Verfahren, da die Strukturen dieser Methoden weitgehend auf den gleichen Überlegungen beruhen. Für eine weitere Beschäftigung mit biometrischen Methoden sind die beiden Lehrbücher von Lothar Sachs (2004/2018) und Rolf J. Lorenz (1992) sehr zu empfehlen, zumal diese ihrerseits viele Beispiele und zahlreiche Hinweise auf spezielle Fragestellungen enthalten.

Einige weitere, speziellere Methoden, die bei vielen Fragestellungen in der medizinischen Forschung immer wieder von Bedeutung sind, sind Gegenstand des nächsten Kapitels.

Kapitel 6:
Spezielle Verfahren

Kapitel 6 hat verschiedene speziellere Methoden zum Gegenstand, die im täglichen Leben der klinischen Forschung und Praxis häufig zum Tragen kommen. Ein logischer Zusammenhang dieser Methoden ist nur bedingt gegeben, so dass auch jeder Abschnitt für sich gelesen werden kann.

6.1 Qualitätssicherung im Labor

In allen Bereichen eines klinischen Labors, in der radiologischen Diagnostik, überall da, wo Messungen mit technischen Geräten zur täglichen Routine gehören, ist eine ständige Überprüfung der Messergebnisse auf *Richtigkeit* und *Genauigkeit* erforderlich. Dies ist speziell im Labor eine Routinemaßnahme, die zum Vergleich mehrerer Labors auch in Form von sogenannten *Ringversuchen* durchgeführt wird.

Eine Routinemethode der Qualitätssicherung verwendet *Kontrollseren*, die die zu analysierenden Serumbestandteile enthalten. Dabei ist von Seiten der Hersteller genau bekannt, welche Anteile dies sind und welche Variabilität (Messfehler, Messungenauigkeit) bei den Messungen vorliegt. Das Kontrollserum wird nun routinemäßig an z.B. 21 Tagen neben den Patientenseren mitbestimmt. Nach Abschluss des Kontrollzeitraums kann man die gefundenen Resultate mit den Vorgaben des Herstellers vergleichen.

In der Praxis verwendet man auch heute noch sogenannte *Kontrollkarten*. In diese Kontrollkarten wird horizontal die Zeit bzw. das Datum eingetragen, in Y-Richtung der Wert der Kontrollserum-Bestimmung. In der Regel finden sich auf der Kontrollkarte fünf horizontale Linien: Eine für den vom Hersteller angegebenen Mittelwert ("Richtwert") des Kontrollserums und, als Warn- bzw. Kontrollgrenzen, vier weitere für die Bereiche $\mu \pm 2\sigma$ und $\mu \pm 3\sigma$. Üblicherweise wird eine Messreihe verworfen, wenn der zugehörige Kontrollwert die $(\mu \pm 3\sigma)$-Grenzen verlässt; engere oder auch weitere Grenzen sind problemabhängig natürlich ebenfalls denkbar.

Nach den Richtlinien der Bundesärztekammer ("RiLi-BÄK") wird der errechnete Durchschnitt der Messungen mit dem *Richtwert* des Kontrollserums verglichen: Die Differenz ist ein Maß für die *Richtigkeit* und wird als *systematische Messabweichung* bezeichnet. Die Standardabweichung ist ein Maß für die *Präzision* der Messungen. Der *QUAMM* ist der Mittelwert der Messabweichungen um den Richtwert, dies ist in Abbildung 24 in einem Beispiel (Daten aus Sysmex Xtra 2/2007, pp. 1-16) mit dem Symbol "Z" bezeichnet. Schließlich werden mit Hilfe eines laborintern modifizierten QUAMM die internen Fehlergrenzen ("F" in Abbildung 24) des

Labors ermittelt, die gewisse vorgegebene Grenzwerte nicht überschreiten dürfen. Einzelheiten und weitere Parameter wie *Relativer QUAMM*, *Relative Messabweichung* und *Variationskoeffizient* mit deren kritischen Grenzwerten finden sich in den aktuellen Richtlinien der Bundesärztekammer.

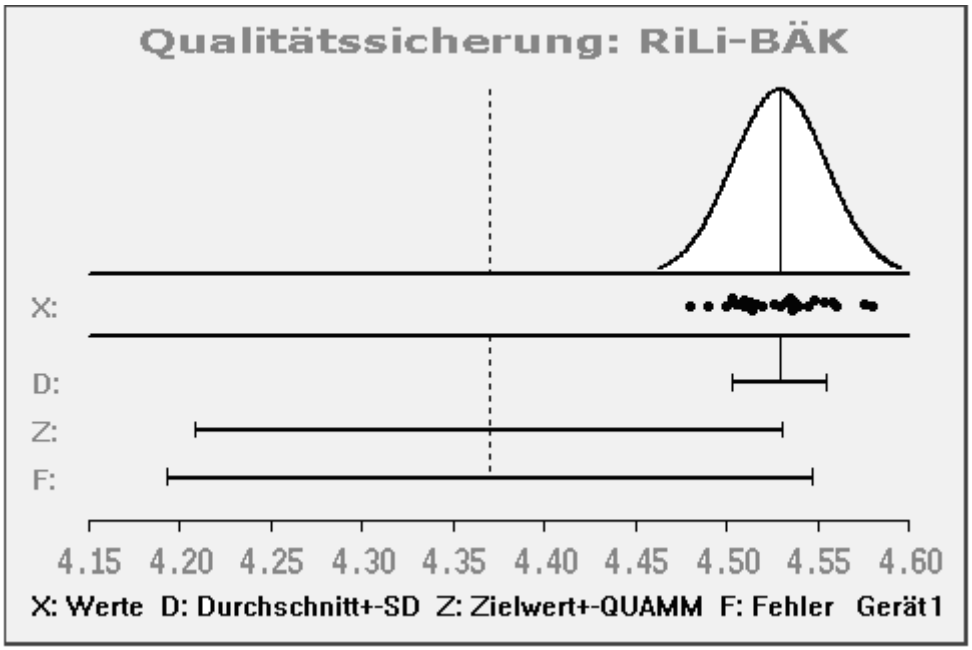


Abbildung 24: Qualitätssicherung gemäß BÄK-Richtlinien

Bei vielen weiteren Fragestellungen, zum Beispiel zur Feststellung des Anteils fehlerhafter Analysen pro Untersuchungs- oder Zeiteinheit, aber auch bei qualitativen Analysen werden ebenfalls Kontrollkarten verwendet. Diese basieren in der Regel auf der *Binomial-, Poisson-* oder auf der *Hypergeometrischen Verteilung*, was aber hier nicht weiter ausgeführt wird. Eine Anwendung ergibt sich zum Beispiel bei der Untersuchung des Anteils fehlerhafter Röntgenaufnahmen pro Tag oder bei der Überprüfung der Materiallieferungen der Hersteller von Labor- oder Röntgenmaterial.

6.2 Bland-Altman-Methodenvergleich

In Abschnitt 3.7 ("Scattergram") wurde in Abbildung 11 ein Beispiel zum Vergleich von zwei Messgeräten vorgestellt. Die Fragestellung lautete, ob ein Messgerät, das vielleicht sehr präzise arbeitet, aber teuer im Unterhalt oder bedienungsunfreundlich ist, durch ein anderes Messgerät ersetzt werden kann, das möglicherweise erfreulichere Eigenschaften besitzt als das bisher verwendete Gerät. Analoges gilt für den Vergleich zum Beispiel von Analysemethoden oder Ähnlichem.

Ein erster Vergleich der beiden Messgeräte wurde in Abschnitt 5.7 mit Hilfe der Korrelationsrechnung vorgenommen. In diesem Zusammenhang wurde bereits festgestellt, dass eine Korrelation der beiden Messungen natürlich unbedingt erforderlich ist, andererseits aber aus dem Vorhandensein einer Korrelation noch kein weiterer Schluss gezogen werden kann: Ein zufriedenstellendes Ergebnis liegt sicher erst dann vor, wenn die Punktwolke "dicht" um die Winkelhalbierende im Scattergram angeordnet ist und damit auf eine mögliche "Identität" der beiden Methoden hinweist.

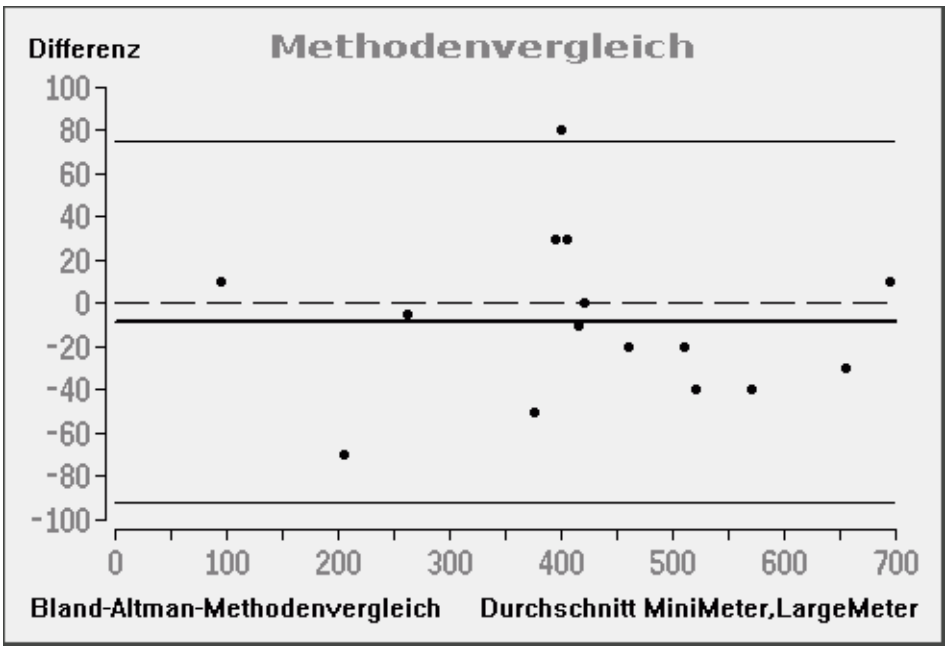


Abbildung 25: Methodenvergleich: Differenz vs. Durchschnitt

Ein einfacher Ansatz zur Bewältigung dieses Problems wurde von J.M. Bland und D.G. Altman 1986 in der Fachzeitschrift "Lancet" vorgestellt. Bezeichnet man die n=15 Messpaare wieder mit (x_i,y_i) und berechnet jeweils die Differenz $\delta_i=x_i-y_i$ und den Durchschnitt $d_i=(x_i+y_i)/2$, so kann man diese beiden neuen Größen δ_i und d_i in einer Graphik gegeneinander auftragen, wie in Abbildung 25 gezeigt wird.

Bei Messgrößen unterschiedlicher Dimension (bei "inkommensurablen Größen") kann man an eine Standardisierung $x'=(x-\overline{x})/s_x$ bzw. $y'=(y-\overline{y})/s_y$ der beiden Messgrößen x und y denken, um Bland und Altman's Methode anwendbar zu machen. Die Interpretation der δ' und d' kann analog zu der hier diskutierten Darstellung erfolgen:

Die Größe der Differenz δ darf nicht von der Größe des Durchschnittes d abhängen; dies überprüft man via Regressionsrechnung (hier: b=0.01, p=0.86, $p>\alpha=0.05$). Weiterhin sollte die mittlere Differenz δ_M idealerweise Null betragen (hier: $\delta_M=-8.3$), gewisse "zufällige" Abweichungen von Null sind zu tolerieren, was mit dem t-Test überprüft werden kann.

Sicherlich spielt bei der Beurteilung von δ_M die Standardabweichung s_δ der Differenzen eine Rolle. In Anlehnung an die genannten Autoren kann man mit Hilfe von δ_M und s_δ ein Konfidenzintervall für die "wahre" mittlere Abweichung μ_δ der Messpaare berechnen; dieses ergibt sich im Beispiel bei einer Konfidenz von P=0.95 mit den Grenzen (-29.2;12.5). Man hat zwar bereits bei Betrachtung von Abbildung 25 und erst recht wegen des Wertes δ_M=-8.3 den Eindruck, dass die Differenzen eher zu negativen Werten neigen, kann diesen Verdacht aber via Nullhypothesentest formal nicht bestätigen.

Zur alternativen Beurteilung wurde ein parametrischer Toleranzbereich für die "übliche" Abweichung der beiden Geräte berechnet (Formel und Erläuterung finden sich im nächsten Abschnitt 6.3); dieser besitzt im Beispiel die in Abbildung 25 dargestellten 95%-Konfidenzgrenzen (-91.6;74.9). Nach Ansicht des Autors liegt damit eine sachgerechte Beurteilungsgrundlage vor: Sind Abweichungen in dieser "üblichen" Größenordnung aus der Sicht des Labors noch zu tolerieren, oder ist man nicht bereit, solche Abweichungen zu akzeptieren? Gemessen an der Gesamtspannweite R=600 der Werte sicher nicht, die Länge des Toleranzbereichs entspricht immerhin 25% der Spannweite. Zu dieser Entscheidung gelangt man damit nicht auf Grund des Scattergrams (Abbildung 11) und erst recht nicht auf Grund des statistisch hoch signifikanten Korrelationskoeffizienten ($p<10^{-6}$!), so dass sich die etwas ausführlichere Analyse - zum Beispiel mit dem Programm **BiAS.** - doch lohnt.

6.3 Normbereiche

Norm- oder Referenzbereiche (auch: *Toleranzbereiche*) finden sich in allen Bereichen der Medizin, dies sowohl unter diagnostischen als auch häufig unter mehr technischen Aspekten. Letztere sollen mit Hinweis auf Abschnitt 6.1 (Qualitätssicherung) nicht weiter vertieft werden, während den diagnostischen Gesichtspunkten etwas mehr Aufmerksamkeit gewidmet werden muss. Mehr dazu finden sich etwa in Ackermann (1994).

"Normbereich" oder "Normalbereich" ist die vielleicht häufiger gewählte, "Referenzbereich" die modernere Bezeichnung für das, womit man gerne Zugang zu einer Definition von "normal" oder "gesund" finden möchte. Eine solche Definition von "gesund" ist eher problematisch und soll auch hier nicht versucht werden, sondern es soll ein formaler Weg zur Berechnung von "Referenzbereichen" gezeigt werden. Dazu sind zunächst einige Definitionen erforderlich:

Ein *Referenz-Individuum* ist ein Individuum, das nach fest definierten *Referenzkriterien* ausgewählt wird. Die *Referenz-Population* besteht aus allen denkbaren Referenz-Individuen, von denen man Referenz-Werte

bestimmen kann. Eine *Referenz-Stichprobe* ist eine der Referenz-Population - idealerweise - nach Zufallskriterien entnommene, repräsentative Stichprobe. Aus den Stichprobenwerten wird mit geeigneten statistischen Mitteln ein *Referenz-Bereich* berechnet, wozu wieder, je nach Verteilungsform der Referenz-Stichprobe bzw. -Population, parametrische oder auch nicht-parametrische Methoden verwendet werden können. (Vergleichen Sie dazu bitte auch die analoge Definition von "Stichprobe und Grundgesamtheit" in Abschnitt 2.1!)

Zur Berechnung von Referenz- bzw. Normbereichen stehen wie auch in anderen Teilgebieten der Statistik parametrische und nicht-parametrische Verfahren zur Verfügung. Parametrische Normbereiche unterstellen wie andere parametrischen Verfahren als grundlegende Voraussetzung eine Gauß-Verteilung der gemessenen Werte und berechnen sich nach der Formel

$$\left[\bar{x} - t_{P,n-1} \cdot s \cdot \sqrt{\frac{n+1}{n}} \;\; , \;\; \bar{x} + t_{P,n-1} \cdot s \cdot \sqrt{\frac{n+1}{n}} \right]$$

P heißt jetzt *Überdeckung* des Normbereichs, denn der Normbereich sollte ja einen Anteil von P - fast immer verwendet man P=95% - der unterstellten Grundgesamtheit aller "Gesunden" erfassen. (Vorsicht: Die Überdeckung P des Normbereichs ist nicht zu verwechseln mit der Konfidenz P des Konfidenzintervalls, dieses bezieht sich auf den Erwartungswert μ und nicht auf die Einzelwerte!!) Vorausgesetzt ist dabei eine eingipflig-symmetrische Verteilung der Referenz-Messwerte. Im Zusammenhang mit Konfidenzintervall und t-Test konnte diese Voraussetzung durch den Zentralen Grenzwertsatz etwas abgeschwächt werden, da dort im Grunde nur die Verteilung der Durchschnittswerte betrachtet wurde, hier dagegen ist man in aller Strenge auf eine Einhaltung der Voraussetzung "Gauß-Verteilung der Einzelwerte" angewiesen. Es gibt nicht viele Beispiele, die dies ruhigen Gewissens zulassen, eine Ausnahme ist zum Beispiel die Verteilung von Leukozytenzahlen in Abbildung 26 (konstruiert nach Angaben aus den Wissenschaftlichen Tabellen Geigy):

Die parametrischen 95%-Normgrenzen der Leukozytenwerte liegen bei [2800;11200]. Von einer Angabe der beliebten "$\bar{x} \pm 2s$"-Grenzen als Normbereich ist prinzipiell abzuraten, denn erstens gewährleisten diese keineswegs eine Überdeckung P (z.B. P=95%) und lassen darüber hinaus viele Anwender eine Untersuchung der Stichprobenwerte auf Eingipfligkeit und Symmetrie gerne vergessen.

Eine sinnvollere Alternative stellen nicht-parametrische Normbereiche oder auch Percentilen ("Quantilen", dazu auch Abschnitt 3.2) dar, die bereits in Abschnitt 3.2 besprochen wurden. Diese können völlig voraussetzungsfrei berechnet werden, führen bei symmetrisch verteilten Daten im Wesentlichen zu den gleichen Ergebnissen wie die parametrischen Methoden und sind insbesondere auch für schiefe Verteilungen berechen-

bar. Da aber nun schiefe Verteilungen, wie bereits weiter oben begründet wurde, in der Medizin und Biologie nicht die Ausnahme, sondern eher die Regel darstellen, gibt es keinen Grund, parametrische Bereiche oder vielleicht "$\bar{x} \pm 2s$"-Grenzen zu verwenden. Percentilen ("Quantilen") bzw. nicht-parametrische Normbereiche sind die Methode der Wahl.

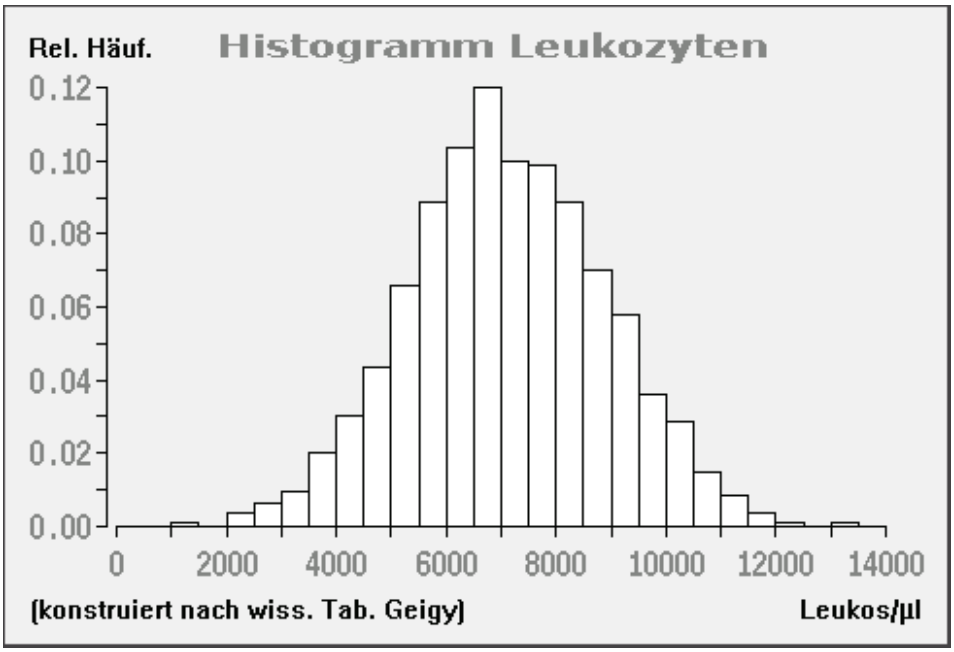


Abbildung 26: Verteilung von Leukozytenwerten

Im Beispiel der Körpergrößen von Neugeborenen (Daten in Kapitel 3, Abbildung 7) errechnet sich der nicht-parametrische 95%-Normbereich mit den Grenzen [40.8 ; 57.5]. Mit den Körpergrößen liegt ein seltenes Beispiel vor, das auch eine Berechnung von parametrischen Grenzen zugelassen hätte; diese liegen bei [41.2 ; 58.7] und weichen nur unwesentlich von den nicht-parametrischen Bereichsgrenzen ab.

Wenn ein "Gesunder" mit seinem Laborwert o.ä. im Norm- bzw. Toleranzbereich liegt und deshalb als "gesund" bezeichnet wird, so spricht man von einer *richtig-negativen* Entscheidung. Liegt der "Gesunde" jedoch außerhalb des Normbereichs (und das gilt bei zum Beispiel P=95% per definitionem in 5% aller Fälle!), so ist er *falsch-positiv*, weil man ihn als "krank" bezeichnet, obwohl er in Wirklichkeit gesund ist.

Den "Kranken" ergeht es gerade umgekehrt: Liegt ein "Kranker" im Normbereich, so ist er *falsch-negativ* und wird somit fälschlich als "gesund" bezeichnet, liegt er aber außerhalb des Normbereiches, so ist er *richtig-positiv* und wird damit korrekterweise als "krank" erkannt.

Im nächsten Abschnitt werden die genannten und einige weitere Definitionen im allgemeineren Rahmen einer "Bewertung diagnostischer Tests" ausführlicher besprochen.

6.4 Bewertung diagnostischer Tests

Zur Bewertung von diagnostischen Tests werden einige charakteristische Begriffe verwendet, die im Folgenden definiert werden:

Die Bezeichnungen (rp), (fn), (fp) und (rn) in den vier Feldern der Tabelle 14 sind Abkürzungen für richtig/falsch negativ/positiv und als Definitionen bereits aus dem letzten Abschnitt bekannt. Diese Definitionen bilden die Grundlage für alle weiteren, in der aktuellen Terminologie üblicherweise verwendeten Begriffe.

Diagnose	Testergebnis positiv	Testergebnis negativ
krank	**richtig positiv (rp)**	**falsch negativ (fn)**
gesund	**falsch positiv (fp)**	**richtig negativ (rn)**

Tabelle 14: Zur Bewertung diagnostischer Tests

Wieviele Kranke werden von einem diagnostischen Test auch als "krank" erkannt? Die *Sensitivität* ist definiert als die Rate "richtig positiv" und wird komplementiert durch die Rate "falsch negativ".

Wieviele Gesunde werden von einem diagnostischen Test als "gesund" bestätigt? Die *Spezifität* ist definiert als die Rate "richtig negativ" und wird komplementiert durch die Rate "falsch positiv".

Bitte beachten Sie den formalen Zusammenhang mit dem Testen von Nullhypothesen und insbesondere Abbildung 18 in Abschnitt 5.1: Die Nullhypothese lautet "Der Patient ist gesund". Damit entspricht einer richtig negativen Entscheidung die korrekte Beibehaltung der Nullhypothese und "richtig positiv" der korrekten Ablehnung von H_0, während "falsch positiv" bzw. "falsch negativ" dem bekannten Fehler 1. bzw. 2. Art entsprechen.

Den nachfolgend definierten *Prädiktiven Werten* liegt Bayes' Formel aus 0.4 zugrunde:

Wieviele Positive sind wirklich positiv? Der *Prädiktive Wert Positiv* ist der Anteil der tatsächlich Kranken bezogen auf alle positiv Getesteten, also der Quotient von "Anzahl richtig positiv" und "Anzahl positiv gesamt".

Wieviele Negative sind wirklich negativ? Der *Prädiktive Wert Negativ* ist der Anteil der tatsächlich Gesunden bezogen auf alle negativ Getesteten als Quotient von "Anzahl richtig negativ" und "Anzahl negativ gesamt".

Wieviele Diagnosen sind richtig? Die *Effizienz* ist die Rate der richtigen Entscheidungen als der Quotient "Anzahl 'richtig positiv' plus Anzahl 'richtig negativ'" dividiert durch "Anzahl gesamt".

Alle Berechnungen sind im Grunde nur einfache Prozentberechnungen mit unterschiedlichen Bezugsgrößen. Es muss unbedingt beachtet werden, dass diese Prozentzahlen im statistischen Sinn nur Schätzungen und insbesondere nur Punktschätzungen darstellen, somit also keine Sicherheiten im Sinne von Konfidenzintervallen implizieren. Bei "kleinen" Fallzahlen können durchaus noch erhebliche Ungenauigkeiten (vgl. Abschnitt 4.1, richtig und genau!) auftreten. Das Programm **BiAS.** gibt zu allen Schätzwerten auch Konfidenzintervalle an, darüber hinaus auch weitere relevante Kenngrößen wie z.B. den *Youden-Index* und die *Likelihood-Ratio*s.

In der Regel ist es zweckmäßig, etwa gleich viele Erkrankte wie Gesunde zu untersuchen, um eine gewisse Vergleichbarkeit der Schätzgrößen zu gewährleisten. Zur Berechnung der beiden Prädiktiven Werte und der Effizienz ist eine repräsentative Stichprobe bzw. Kenntnis der Prävalenz vorausgesetzt; in **BiAS.** ist optional eine Vorgabe der Prävalenz möglich.

Zum Nachrechnen findet sich in Abbildung 27 ein einfaches Zahlenbeispiel: Für rp=87, fn=2, fp=13 und rn=100 erhält man mit dem Programm **BiAS.** das Ergebnis inclusive 95%-Konfidenzintervalle (in Klammern):

Bewertung diagnostischer Tests

```
                  |         Testergebnis:          |  Summe der
                  |    positiv    |    negativ     |      Zeile
------------------+---------------+----------------+-----------
Anzahl Erkrankte  |   87 97.75%   |    2  2.25%    |        89
------------------+---------------+----------------+-----------
Anzahl Gesunde    |   13 11.50%   |  100 88.50%    |       113
------------------+---------------+----------------+-----------
Summe der Spalte  |  100          |  102           |       202

Prävalenz (errechnet, Anteil erkrankt/gesamt) = 44.1%(37.1%-51.2%)
Sensitivität (richtig positiv) =       87/ 89 = 97.8%(92.1%-99.7%)
Spezifität   (richtig negativ) =      100/113 = 88.5%(81.1%-93.7%)
Rate 'falsch positiv' =                13/113 = 11.5%( 6.3%-18.9%)
Rate 'falsch negativ' =                 2/ 89 =  2.2%( 0.3%- 7.9%)
Prädiktiver Wert positiv (ri.pos./pos.gesamt) = 87.0%(78.8%-92.9%)
Prädiktiver Wert negativ (ri.neg./neg.gesamt) = 98.0%(93.1%-99.8%)
Effizienz (Rate der richtigen Entscheidungen) = 92.6%(88.1%-95.8%)
Youden-Index Y = Sensitivität+Spezifität-100% = 86.3%(79.6%-92.9%)
Likelihood-Ratio positiv = Sens./(1-Spez.)    =  8.50 (5.09-14.18)
Likelihood-Ratio negativ = (1-Sens.)/Spez.    =  0.03 (0.01- 0.10)
```

Abbildung 27: Programmausgabe zur Bewertung von diagnostischen Tests

Leserinnen und Leser, die mehr über das Thema wissen möchten, können auf das ausgezeichnete und sehr umfassende Buch von Ulrich Abel, erschienen im Hippokrates-Verlag 1993, verwiesen werden.

6.5 ROC-Analyse

Bei der Konstruktion eines diagnostischen Tests muss man häufig auf Grundlage von stetigen Größen, zum Beispiel von Laborwerten, einen *Schwellenwert* definieren, bei dessen Über- oder Unterschreiten auf das Vorliegen einer Erkrankung geschlossen wird. ROC-Kurven ("receiver-operating characteristic") liefern einen visuellen Eindruck von der "Güte" des Diagnostischen Tests, wogegen sich Abschnitt 6.6 mit der Berechnung einer "optimalen" Schwelle befasst.

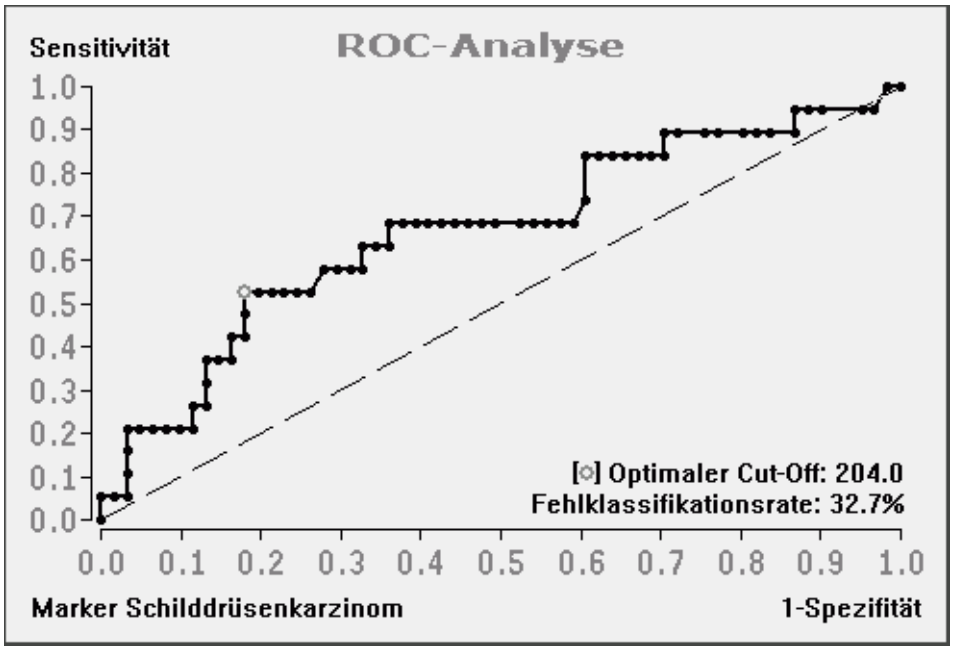


Abbildung 28: ROC-Analyse

In der Schilddrüsendiagnostik wird ein neuer Tumormarker untersucht. Dazu werden die Daten von n_1=19 erkrankten Personen mit n_2=69 Gesunden verglichen. Die n=19+69=80 Werte werden in eine gemeinsame Rangfolge gebracht und der Reihe nach alle möglichen "Trennpunkte" T betrachtet: Der i-te Trennpunkt T_i liegt zwischen dem i-ten und dem (i+1)-ten Wert in der gemeinsamen Rangfolge. Zu jedem möglichen Trennpunkt T_i werden die Sensitivität und die Spezifität berechnet (dazu Abschnitt 6.4) und die ermittelten Werte in Abbildung 28 dargestellt (es ist Sensitivität = richtig positiv und 1-Spezifität = falsch positiv, Likelihood-Ratio = Sensitivität / (1-Spezifität), vgl. Abbildung 27!).

Fällt die Kurve mit der Winkelhalbierenden überein, so besitzt der Test keine diagnostische Aussagekraft: Die Fläche unter der Kurve ("AUC") kann als Maß für die "Güte" des Tests verwendet werden. Statistische Tests findet man bei Abel (1993), **BiAS.** führt die Berechnungen durch.

6.6 Diskriminanzanalyse

In einem Labor wird ein neues Isoenzym der LDH entdeckt. Erste Untersuchungen sprechen dafür, dieses Isoenzym in der Infarktdiagnostik einzusetzen, da es bei Infarktpatienten im Mittel höhere Werte aufzuweisen scheint als bei Patienten ohne akuten Infarkt (modif. aus Werner (1992)).

Um diese Vermutung zu überprüfen, wird eine Pilotstudie mit $n_1=20$ Infarktpatienten und einer Kontrollgruppe mit ebenfalls $n_2=20$ Personen ohne Infarkt durchgeführt. Da die gemessenen Werte erwartungsgemäß rechtsschief verteilt sind, wird zwecks Symmetrisierung der Verteilungen (Gauß-Verteilung!) eine Log-Transformation der Daten durchgeführt. Es ergeben sich hieraus die Werte $\overline{x}_1=7.72$, $\overline{x}_2=8.72$, $s_1=0.69$ und $s_2=0.71$. Der F-Test zur Überprüfung der Gleichheit der Varianzen ist negativ, so dass ein Zweistichproben-t-Test zur Prüfung der Nullhypothese $H_0(\mu_1=\mu_2)$ durchführbar ist. Dieser lehnt die Nullhypothese am Signifikanzniveau $\alpha=0.01$ ab, da der errechnete t-Wert mit $t=4.52$ deutlich über dem kritischen Wert $t_{0.99,38}=2.71$ liegt; die Überschreitungswahrscheinlichkeit beträgt sogar $p=0.00006$: Es besteht ein "statistisch hoch signifikanter" Unterschied zwischen der Infarkt- und der Kontrollgruppe bezüglich der mittleren (!) Enzymkonzentrationen der beiden Gruppen. Kann damit das neue Isoenzym Einzug in die Diagnostik halten?

Die Diskriminanzanalyse befasst sich in Hinblick auf die Einzelwerte mit dem Problem der "optimalen" Trennung zweier oder mehrerer Gruppen anhand von im Allgemeinen mehreren medizinischen "Parametern". Im oben skizzierten Beispiel liegt nur eine Variable vor; die grundlegenden Prinzipien der Diskriminanzanalyse lassen sich dabei jedoch sehr einfach darstellen:

In Hinblick auf einen diagnostischen Einsatz des Isoenzyms ist es wünschenswert, die beiden Gruppen so zu trennen, dass man von einer minimalen Rate von Fehldiagnosen ausgehen darf. Im vorliegenden Fall im Wesentlichen gleicher Streuungen errechnet sich der gewünschte "optimale Trennpunkt" als gewichteter Durchschnitt der beiden Durchschnitte, hier bei gleichen Stichprobenumfängen ganz einfach durch $x_T=(\overline{x}_1+\overline{x}_2)/2$. Als Zuordnungsregel definiert man

$x < x_T \Rightarrow$ Kein Infarkt
$x > x_T \Rightarrow$ Infarkt

Man kann mathematisch zeigen, dass der Trennpunkt x_T in der oben geforderten Weise "optimal" ist, also eine minimale Rate falsch-positiver und falsch-negativer Entscheidungen erwarten lässt. Diese "Fehlklassifikationsrate" kann im vorliegenden Spezialfall über die Standardisierung

$$\varphi_i = \frac{x_T - \bar{x}_i}{\sqrt{\frac{(n_1 - 1)\cdot s_1^2 + (n_2 - 1)\cdot s_2^2}{n_1 + n_2 - 2}}}$$

anhand der standardisierten Gauß-Verteilung ermittelt werden (der Nenner ist die gewichtete mittlere Standardabweichung der beiden Stichproben): Im Beispiel ist $\varphi_1=\varphi_2=1.4286$, die geschätzte Rate der beiden möglichen Fehldiagnosen "falsch-negativ" und "falsch-positiv" beträgt somit $2\cdot 0.0766=0.1531=15.31\%$. (Der Wert $p=0.0766$ ergibt sich bei *ein*seitiger Betrachtung für $\varphi=1.4286$ durch Interpolation aus Tabelle 4 bzw. exakt mit Hilfe eines Programms, z.B. **BiAS**.) Die Entscheidung, ob dieses Ergebnis - möglicherweise im "Widerspruch" zur statistischen Signifikanz des Vergleichs der Mittelwerte via $H_0(\mu_1=\mu_2)$! - noch als medizinisch relevant angesehen werden darf, muss der Entscheidung des diagnostizierenden Arztes überlassen bleiben.

Abbildung 29 zeigt die wesentlichen Definitionen zur Interpretation der Diskriminanzanalyse:

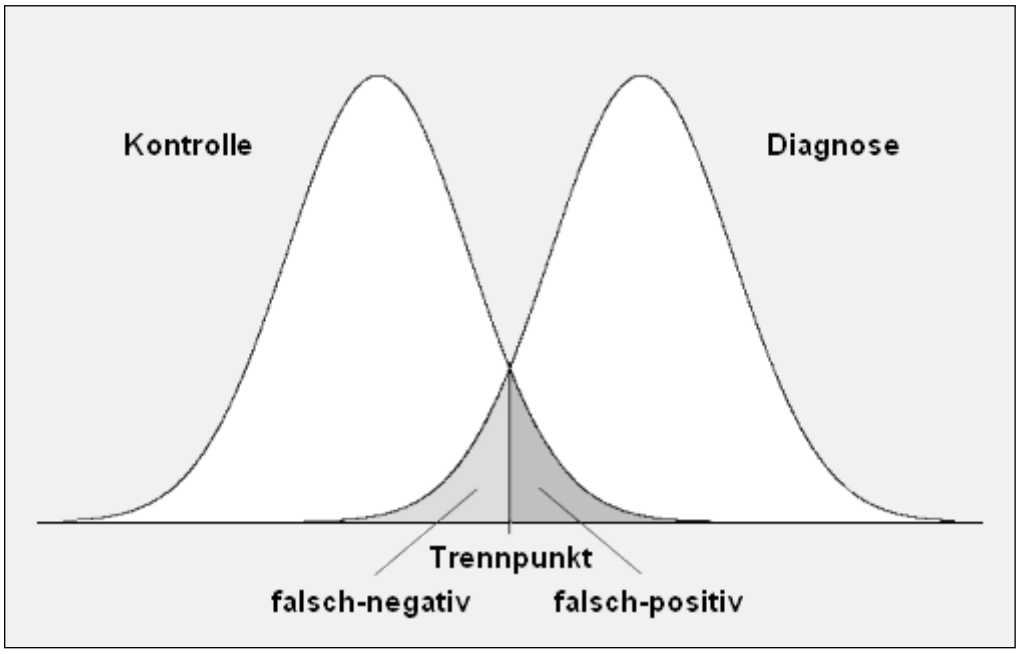


Abbildung 29: Diskriminanzanalyse

Da nun kein Mediziner auf Grund von nur einem "Parameter" – im Beispiel das Isoenzym HBDH - eine Diagnose stellen wird, bietet sich bei Berücksichtigung mehrerer Parameter eher die Durchführung einer multivariaten, also mehrdimensionalen Diskriminanzanalyse mit mehreren Variablen an. Eine Darstellung führt über den Rahmen dieses einführenden Textes hinaus, so dass interessierte Leserinnen und Leser an dieser Stelle auf das

sehr aufschlussreiche und kompetent geschriebene Buch von Deichsel und Trampisch, "Clusteranalyse und Diskriminanzanalyse" verwiesen werden.

Wie bei statistischen Testverfahren stehen auch zur Diskriminanzanalyse nicht-parametrische Methoden zur Verfügung, wozu aber ebenfalls auf die einschlägige Literatur verwiesen werden muss. Das Programm **BiAS.** stellt Berechnungsmethoden für alle genannten Fragestellungen zur Verfügung.

6.7 Überlebenszeitanalyse

Die Bezeichnung *Überlebenszeitanalyse* (englisch: "*Survival-Analysis*") ist historisch bedingt, da die entsprechenden Methoden in der Analyse von eben Überlebenszeiten von Karzinompatienten ihre erste Rolle spielten. Allgemeiner spricht man auch von einer sogenannten *Time-to-Event-Analysis*, denn man kann mit diesen Verfahren alle *Zeit-bis-Ereignis-Fragestellungen* untersuchen: Das sind also nicht nur Fragen, die sich mit Überlebenszeiten beschäftigen, es kann auch die Haltbarkeit von Gelenksprothesen untersucht werden, die Zeit, bis ein Kontrollserum die vorgegebenen Bestimmungsgrenzen verlässt oder die Zeit bis Heilung einer Erkrankung, die Zeit also, an der ein bestimmtes *Zielereignis* eintritt. Irgendwann möchte man die bisher vorliegenden Ergebnisse auswerten, und fast immer gibt es Merkmalsträger, bei denen das Zielereignis zum Auswertungszeitpunkt noch nicht eingetreten ist. Zum Beispiel möchte man eine onkologische Therapiestudie zu einem bestimmten, im Versuchsplan vorgesehenen Zeitpunkt auswerten, und zu diesem Zeitpunkt ist das Zielereignis "Tod" (oder "Rezidiv") bei vielen Patienten noch nicht eingetreten. Es ist deshalb sinnlos, einen "gewöhnlichen" Mittelwert wie in Abschnitt 3.1 oder den Median zu berechnen, denn, was ersteren betrifft, sind Überlebenszeiten ohnehin nicht symmetrisch verteilt, und andererseits ist natürlich nicht bekannt, wie lange einzelne Patienten noch leben werden, so dass man mit einem "Mittelwert" sicher zu kleine, "verzerrte" Werte erhalten wird.

Eine etablierte Methode zur Behandlung von Zeit-bis-Variablen ist die nach den Autoren benannte *Kaplan-Meier-Methode*. Diese berücksichtigt auch sogenannte *zensierte Beobachtungen*, also die Tatsache, dass im Beispiel zum Auswertungszeitpunkt Patienten noch leben und damit deren Überlebenszeit noch nicht feststeht, andere dagegen bereits verstorben sind; nur wenn wirklich alle Patienten gestorben wären, würde die Errechnung des Medians sinnvoll sein. Zensierte Beobachtungen liegen auch dann vor, wenn ein Patient zum Beispiel aus anderen, nicht mit der Untersuchung zusammenhängenden Gründen verstorben ist (Autounfall o.ä.) oder vielleicht ganz einfach verzogen ist und sich damit der weiteren Beobachtung entzieht: Bis zu diesem Zeitpunkt lebte der Patient noch, aber wie lange er noch lebt, weiß man nicht.

Von der Idee her berechnet die Kaplan-Meier-Methode zu jedem Todeszeitpunkt t_i eines Patienten den Quotienten Q_i = (Anzahl zur Zeit t_i nachweislich noch lebender Patienten) / (Anzahl im *vollständigen* Intervall t_{i-1} bis t_i nachweislich "unter Risiko" stehender Patienten) als Schätzung für die Wahrscheinlichkeit, dass ein Patient Zeitpunkt t_i überlebt, vorausgesetzt, er überlebte Zeitpunkt t_{i-1} ("*bedingte Wahrscheinlichkeit*", Abschnitt 0.3). Die eigentlich interessierende Wahrscheinlichkeit für einen zukünftigen Patienten, Zeitpunkt t_i zu überleben, schätzt man auf Grund der Stichprobe via $P_i=Q_1 \cdot Q_2 \cdot \ldots \cdot Q_i$ (Rechenregeln für Wahrscheinlichkeiten, Abschnitt 0.2). Offensichtlich ist diese Berechnungskonvention willkürfrei und erfordert keine speziellen Voraussetzungen, insbesondere gehen nach Kaplan und Meier für jedes Q_i – ungeachtet einer vielleicht späteren Zensierung – immer nur diejenigen Patienten in die Berechnung ein, die im *gesamten* Zeitintervall t_{i-1} bis t_i *unter Risiko* (engl.: "*under risk*") standen.

Wenn auch die Berechnung der Kaplan-Meier-Schätzer prima vista nicht allzu schwierig erscheint, ist wegen des Umgangs mit zensierten Werten ein einfaches Beispiel hilfreich (Tabelle 15). Im Beispiel führen zwei Ereignisse dazu, dass zensierte Werte vorliegen: Fünf Patienten kommen zu einer monatlichen Untersuchung ("lebt"), zur nächsten nicht mehr; über das Schicksal der Patienten besteht Unklarheit. Zwei Patienten melden sich zu einem Zeitpunkt beim Studienleiter ab, weil sie in eine andere Stadt umziehen ("verzogen"). Drei der zehn Patienten sind verstorben.

Name	Überlebenszeit	Status
I	53	verstorben
II	60	lebt
III	85	verstorben
IV	135	verzogen
V	150	lebt
VI	150	lebt
VII	195	verstorben
VIII	210	lebt
IX	240	lebt
X	288	verzogen

Tabelle 15: Rechenbeispiel zur Survival-Analyse

Die Interpretation des Beispiels ist nur an denjenigen drei Zeitpunkten von Interesse, an denen Patienten verstorben sind ("*Events*"), also bei 53, 85 und 195 Tagen, trotzdem gehen alle Patienten in die Berechnung ein, da auch die zensierten Zeiten berücksichtigt sind. Die Ergebnisse mit den oben eingeführten Bezeichnungen sind in Tabelle 16 zusammengestellt, eine graphische Umsetzung der Tabelle findet sich in Abbildung 30.

Tag	Geschätzte Überlebenswahrscheinlichkeit für		
	Tagesbeginn	Tag (Q_i)	Ende des Tages (P_i)
53	100%	9/10	100%·9/10=90%
85	90%	7/ 8	90%·7/ 8=79%
195	79%	3/ 4	79%·3/ 4=59%

Tabelle 16: Ergebnisse zu den Daten aus Tabelle 15

Die graphische Darstellung dieser Ergebnisse in Abbildung 30 ist zwar wegen der kleinen Fallzahl und der relativ großen Anzahl zensierter Werte wenig informativ, zeigt aber doch sehr deutlich das Konstruktionsprinzip. Die kleinen senkrechten Striche auf der Kaplan-Meier-Kurve symbolisieren die sieben zensierten Werte; bitte beachten Sie, dass die Kurve gemäß Tabelle 16 nur bei einem "Event" eine Stufe aufweist und folglich ohne eine solche Stufe mit der maximalen (hier zensierten!) Überlebenszeit von 288 Tagen abbricht. Unter den Zeiten befindet sich die "number at risk":

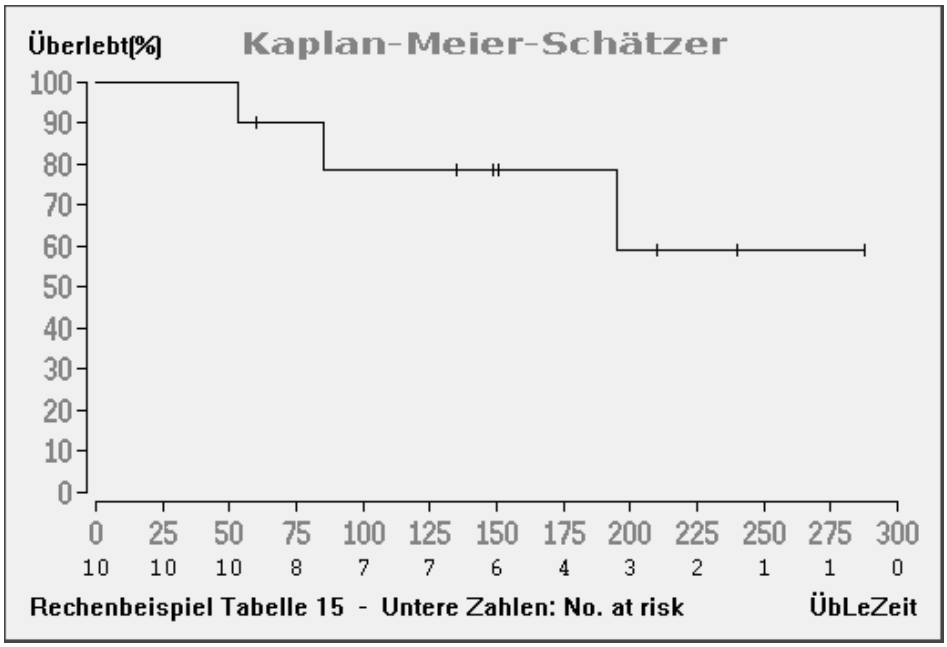


Abbildung 30: Kaplan-Meier-Schätzungen zu den Tabellen 15 und 16

In der Survival-Analyse spielen speziell bei kleinen Stichprobenumfängen wieder Konfidenzintervalle eine maßgebliche Rolle. Die Ergebnisse des letzten Beispiels sind auf Grundlage von nur sehr kleinen Fallzahlen zustande gekommen und lassen einen Anhalt über die statistische "Unsicherheit" der oben geschätzten Überlebenswahrscheinlichkeiten vermissen: In der Tat ergeben sich im Beispiel Konfidenzintervalle, die nahezu den

gesamten Bereich von 0 bis 100% einschließen, so dass auf eine Angabe verzichtet wurde. Stattdessen verschafft das nächste Beispiel auf Grundlage einer deutlich größeren Fallzahl einen Eindruck über die Aussagefähigkeit von Hall-Wellner-Konfidenzbändern, dies jedoch ohne auf die zugehörige Mathematik oder auf Details der Berechnung einzugehen.

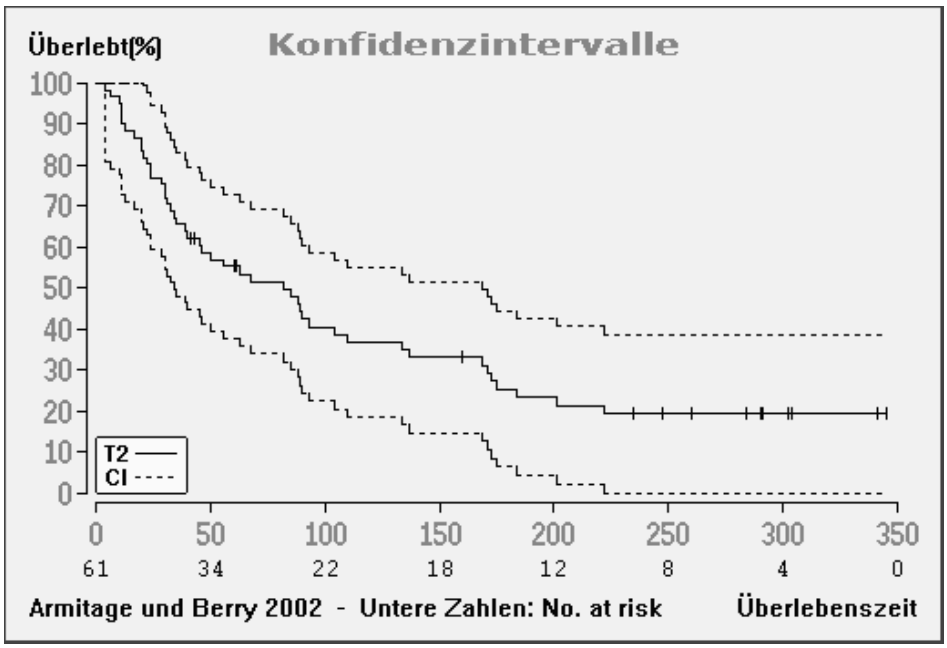


Abbildung 31: Kaplan-Meier-Überlebenszeitschätzungen

Abbildung 31 stellt die Ergebnisse einer onkologischen Überlebenszeitstudie einschließlich der zugehörigen *Hall-Wellner'schen Konfidenzintervalle* graphisch dar. In der Abbildung erkennt man die typisch fallende, treppenförmig verlaufende Kaplan-Meier-Kurve mit den zitierten Konfidenzintervallen für die Überlebenszeiten. Die im Zeitverlauf etwas breiter werdenden Konfidenzbänder (Vorsicht: senkrechte Abstände) sind Ausdruck der ebenfalls im Zeitverlauf wachsenden "Unsicherheit" der Kaplan-Meier-Schätzungen, die durch die geringer werdende Patientenzahl bedingt ist. Der Stichprobenumfang der Studie war n=61 (vgl. Armitage und Berry, Blackwell Publ. 2002).

Bei konkurrierenden Sterberisiken, etwa "Tod am Tumor" und "Tod aus anderen Ursachen", verwendet man günstiger eine Variante des Kaplan-Meier-Schätzers, den *Aalen-Johansen-Schätzer*. Gut lesbare Einzelheiten dazu finden sich in dem Lehrbuch von Held et al. (2013).

Der Vergleich von zwei oder von mehreren Behandlungsgruppen kann mit Hilfe des *Log-Rank-Tests* in verschiedenen Varianten durchgeführt werden. Analog kann für zwei oder mehrere Gruppen der *Relative Hazard* berechnet werden. Kompliziertere Fragestellungen, die auch *Kovariablen* wie Begleiterkrankungen, das Alter der Patienten u.a. berücksichtigen, lassen sich mit dem *stratifizierten Log-Rank-Test* oder optimal mit dem *Cox-Modell* ("*Cox' proportional hazards regression*") bearbeiten: Einzelheiten zu diesen Methoden finden Sie in Abschnitt 6.12 und sehr ausführlich bei Armitage und Berry (2002) oder Held et al. (2013).

6.8 Das Intention-to-Treat-Prinzip

In Kapitel 2 wurden bereits einige Prinzipien der statistischen Versuchsplanung beschrieben, die die Vergleichbarkeit von zwei oder auch von mehreren Studienarmen gewährleisten sollten - man erinnere sich dazu insbesondere an die Randomisierung. Im Verlauf einer Studie kann aber trotz sorgfältiger Planung die Vergleichbarkeit der Studiengruppen gefährdet werden: Speziell bei Studien mit langer Laufzeit kann es sein, dass Patienten aus der Studie ausscheiden, und dies möglicherweise aus Gründen, die mit dem Behandlungserfolg zusammenhängen können. Als zweites Beispiel wäre es denkbar, dass viele Patienten mit "Novum" wegen Nebenwirkungen die Teilnahme an der Studie abbrechen, während dies bei "Standard" nicht der Fall ist. In beiden Fällen kann es unter Umständen zu einer "Verzerrung" der Studienresultate kommen, wie folgendes Beispiel zeigt (modifiziert nach Wilcox et al. BMJ 1980, pp. 885-8):

Infarktanteil %	Beta-Blocker	Referenz
Abbruch	16%	13%
Kein Abbruch	3%	11%
Gesamt	8%	12%

Tabelle 17: Anteil verstorbener Patienten nach Infarkt

Bei den abbrechenden Patienten mit der Beta-Blocker-Therapie ist der Anteil verstorbener Patienten sehr hoch (16%), bei den Nicht-Abbrechern dagegen vergleichsweise niedrig (3%): Offenbar ergibt sich bei Ignorierung der Studienabbrecher eine deutliche Überschätzung des positiven Behandlungseffektes auch in Bezug auf die Referenz, so dass man den Abbruch der Studie möglicherweise in Zusammenhang mit einer ungünstigen Prognose sehen muss. Natürlich ist unklar, ob die Nebenwirkungen tatsächlich zum Versterben der Patienten beitrugen, trotzdem muss diese Information "irgendwie" bei der Analyse berücksichtigt werden.

Das "*Intention-to-Treat-Prinzip*" verlangt, dass *alle* in die Studie aufgenommenen Patienten bei der Auswertung berücksichtigt werden, womit ein Zusammenhang der gemäß Randomisierung intendierten Therapie und deren Konsequenzen - zum Beispiel der Nebenwirkungen - hergestellt wird. Im Studienplan einer Therapiestudie müssen deshalb zwingend Kriterien dokumentiert werden, wie in der Auswertung mit Studienabbrechern (sog. (*dropouts* oder *withdrawals*) zu verfahren ist. Im Beispiel ist es denkbar, dass, in Modifizierung der Zielgröße, die Studienabbrecher als "Therapieversager" aufgefasst werden. In allen vergleichbaren Fällen kann auch das "*last value carried forward*"-Prinzip (*LVCF*) in Erwägung gezogen werden.

Das „Gegenteil“ einer ITT-Analyse ist die Analyse *per protocol*. Dabei werden bei der Analyse nur Patienten berücksichtigt, die sich prüfplankonform verhalten haben, andere, die wie im Beispiel aus der Studie ausgeschieden sind, die Behandlung gewechselt haben oder aus sonstigen Gründen nicht das Protokoll erfüllen, gehen nicht in die Auswertung ein: Ein beobachteter Therapieeffekt kann dadurch erheblich "verzerrt" werden (Trampisch et al. 2000).

6.9 Die Number-Needed-to-Treat

In einer ophthalmologischen Studie zu Katarakt-Operationen wurden zwei verschiedene Operationstechniken verglichen. Die Ergebnisse der Studie finden sich in Tabelle 18: "Standard" ist die Standardtechnik und "Neue OP" die modifizierte neue Technik, bei der, wie bereits vermutet, weniger Komplikationen auftreten:

n und %	Erfolg		Komplikation	
Standard	78	86.7%	12	13.3%
Neue OP	83	92.2%	7	7.8%

Tabelle 18: Erfolgsraten zweier Katarakt-Operationstechniken

Neben den beiden Komplikationsraten CER=12/(78+12)=13.3% ("*Control Event Rate*") und EER=7/(83+7)=7.8% ("*Experiment Event Rate*") interessiert man sich für die Differenz dieser Raten, die die *Absolute Risikoreduktion* ARR darstellt: Es ist ARR=13.3%-7.8%=5.5%. Einzelheiten zu einem teststatistischen Vergleich der beiden Operationsmethoden finden sich in Kapitel 5.6 zum χ^2-Vierfeldertafel-Test.

Die Absolute Risikoreduktion kann man auch relativ zur Komplikationsrate CER der Standard-OP beurteilen und berechnet somit die *Relative Risikoreduktion* RRR=ARR/CER=5.5%/13.3%=41.4%: Bezogen auf den Standard ist das Komplikationsrisiko um 41.4% (bzw. auf 58.6%!) reduziert.

Gelegentlich berechnet man auch die sogenannten *Odds* ("Chancen") als Verhältnis der Erfolgs- zur Misserfolgsrate: Es sind $Odds_S$=78:12=5.5 und $Odds_N$=83:7=11.9. Der Quotient der Odds ist das *Odds-Ratio* ("Chancenverhältnis") OR=$Odds_N/Odds_S$=11.9/5.5=2.2. Bitte beachten Sie die Zusammenhänge mit bzw. die Unterschiede zum Relativen Risiko und zum Zuschreibbaren Risiko, die in Kapitel 0.3 beschrieben wurden! (Hinweis: Mit P=Erfolgsrate ist die Odds=P/(1-P) oder umgekehrt P= Odds/(1+Odds).)

Eine weitere interessante Größe zur Bewertung von Therapien o.ä. ist die sogenannte *Number-Needed-to-Treat* oder kurz NNT. Dies bedeutet die mittlere Anzahl von Patienten mit der neuen Technik, bei denen sich ein Erfolg mehr einstellt als bei der gleichen Anzahl von Patienten mit Standard. Die NNT ist damit definiert als NNT=1/ARR, im Beispiel ist NNT=1/5.5% = 18.2: Bei 19 Patienten mit der neuen OP profitiert ein Patient mehr als unter "Standard".

Grundsätzlich sollte man zu der Number-Needed-to-Treat und natürlich auch zu den anderen oben beschriebenen Größen auch die entsprechenden Konfidenzintervalle mit angeben, denn insbesondere bei kleinen Fallzahlen können diese Kenngrößen noch erhebliche Unsicherheiten bzw. Abweichungen von den "wahren" Werten aufweisen. Formeln dazu können hier nicht angegeben werden: Das Programmpaket **BiAS.** berechnet die Größen EER, CER, OR, ARR, RRR, NNT und einige mehr einschließlich deren Konfidenzintervalle.

6.10 Multiple Lineare Regression

In Abschnitt 5.7 ("Regressions- und Korrelationsrechnung") wurden die Grundlagen der Einfachen Linearen Regressionsrechnung vorgestellt. Das Ziel war, über den Gaußschen Ansatz der *Methode der kleinsten Quadrate* den Einfluss *einer* Einflussgröße X auf eine Zielgröße Y zu untersuchen, wozu eine Geradengleichung verwendet wurde:

$$Y = c + \beta \cdot X = \beta_0 + \beta_1 \cdot X$$

Die Multiple Lineare Regressionsrechnung verwendet nun nicht nur *eine*, sondern *mehrere* Einflussgrößen X_i (i=1,…N), die ebenfalls mit der Gaußschen Methode der kleinsten Quadrate behandelt werden können:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_N \cdot X_N$$

Gerard E. Dallal (2000,2012) diskutiert dazu ein interessantes Beispiel einer Querschnittsstudie, in dem Y=log(HDL) als abhängige Variable und als unabhängige Variablen ("Prädiktoren") X_i die Größen Alter, BMI, Vitamin C im Blut, systolischer und diastolischer Blutdruck, die Hautfaltendicke und log(Gesamt-Cholesterin) verwendet wird. (Die Logarithmierung dient in diesem Beispiel lediglich der Symmetrisierung der Cholesterin-Werte und ist ansonsten methodisch nicht von Bedeutung.)

Tabelle der Varianzanalyse: Abhängige Variable Y=log(HDL)

	df	Summe der Quadrate	Mittlere Quadrate	F	p
Auf der Regression	8	0.54377	0.06797	6.16	<0.0001
Um die Regression	147	1.62276	0.01104		
Total	155	2.16652			

Multipler Korrelationskoeffizient R = 0.5010
Bestimmtheitsmaß: R^2 = B = 0.2510
Mittelwert der abhängigen Größe: Y = 1.7109

Prüfung der Koeffizienten:

Variable	i	b(i)	StdAbw. s	df	t	p
Konstante	0	1.16448	0.28804	147	4.04	<0.0001
AGE	1	-0.00093	0.00125	147	-0.74	0.4602
BMI	2	-0.01205	0.00295	147	-4.08	<0.0001
BLC	3	0.05055	0.02215	147	2.28	0.0239
RRSYST	4	-0.00042	0.00044	147	-0.95	0.3436
RRDIAST	5	0.00255	0.00103	147	2.47	0.0147
GLUM	6	-0.00047	0.00019	147	-2.50	0.0135
SKINF	7	0.00147	0.00183	147	0.81	0.4221
LCHOL	8	0.31109	0.10936	147	2.84	0.0051

Tabelle 19: Programmausgabe der Ergebnisse der Multiplen Regression

Liegen für einen künftigen Patienten die Werte $X_i=x_i$ der Prädiktoren Alter, BMI etc. vor, so kann man mit den Koeffizienten $b_i=b(i)$ (den Schätzwerten für β_i aus Tabelle 19!) gemäß der Modellgleichung der Multiplen Regression den Wert von Y=log(HDL) und, per exp(log(HDL)), den zu erwartenden Wert des HDL-Cholesterins berechnen. Wie im Univariaten gibt b_i die erwartete Veränderung von Y an, wenn die Variable X_i um 1 vergrößert wird und alle anderen Prädiktoren konstant bleiben.

Diese Betrachtung gilt für quantitative und für nominale bzw. binäre 0/1-Variablen. Bei ordinalen Daten mit k Stufen kann man eine Dummy-Codierung mit k-1 nominalen Variablen vornehmen, wie im Zusammenhang mit der Cox-Regression in Abschnitt 6.12 gezeigt wird.

Tabelle 19, die man zum Beispiel mit Hilfe des Programmpaketes **BiAS.** berechnen kann, ist weitgehend selbsterklärend. Im ersten Teil wird die Aussagefähigkeit des Modells untersucht (statistisch zu testen anhand des p-Wertes), der zweite Teil befasst sich mit den einzelnen Prädiktoren und deren Stellenwert im Modell:

Der Streuungsanteil "auf der Regression" bedeutet den Anteil der Gesamtstreuung "total", der durch das Modell erklärt werden kann, "um die Regression" bedeutet die Residualstreuung, die nicht durch das Modell erklärt werden kann. Die Mittleren Quadrate bedeuten den Quotienten aus der Summe der Quadrate und dem korrespondierenden Freiheitsgrad.

Via F bzw. p prüft man, ob das Modell als Ganzes einen Erklärungswert für die Zielgröße besitzt. Das Bestimmtheitsmaß B ist analog zur Einfachen Linearen Regression erklärt (vgl. Abschnitt 5.7) und gibt den Anteil der durch das Modell erklärten Summe der Quadrate an.

Die Standardabweichungen s_i beziehen sich auf die einzelnen Regressionskoeffizienten b_i. s_i und b_i kann man zur Konstruktion von Konfidenzintervallen für die Koeffizienten β_i benutzen; die (angenäherten) P·100%-Konfidenzintervalle berechnen sich mit

$$[\ b_i - u_P \cdot s_i\ ;\ b_i + u_P \cdot s_i\ ]$$

wobei u_P die P·100%-Quantile der Gauß-Verteilung darstellt (vgl. dazu Abschnitt 4.2, für z.B. P=0.95 ist $u_{0.95}=1.96$).

Die Teststatistik t prüft die Hypothese $H_0(\beta_i=0)$, dass ein Regressionskoeffizient $\beta_i = 0$ ist *falls alle anderen Prädiktoren im Modell enthalten sind*. (In manchen Programmen wird eine äquivalente Prüfung via F-Verteilung vorgenommen). Eine Beurteilung erfolgt, wie üblich, mit Hilfe der p-Werte.

Es ist zu betonen, dass ein nicht-signifikanter p-Wert lediglich bedeutet, dass bei Anwesenheit der übrigen Parameter die entsprechende Variable keinen *zusätzlichen* prädiktiven Aufschluss bezüglich der Zielgröße beinhaltet, nicht aber, dass die Variable *alleine* betrachtet keinen Erklärungswert für die Zielgröße besitzt.

Die Auffassung, dass man alle nicht-signifikanten Prädiktoren aus dem Modell entfernen kann ist ebenfalls nicht korrekt: Entfernt man einen nicht-signifikanten Prädiktor, so können möglicherweise anschließend andere, bisher nicht-signifikante im neuen reduzierten Modell durchaus signifikant werden.

Aus der Signifikanz des Gesamtmodells kann nicht auf Signifikanz der einzelnen Prädiktoren geschlossen werden, das Umgekehrte ist ebenfalls nicht möglich:

Im Vergleich der beiden Teile der Tabelle 19 kann es zu - scheinbaren - Widersprüchen kommen, die aber in fast allen Fällen einfach erklärt werden können. Dazu im Speziellen:

- Ein mutmaßlich "sicherer" Prädiktor ist nicht signifikant: Sind zum Beispiel Körperlänge und –gewicht im Modell, so kann etwa der Bodymass-Index BMI (als aus Länge und Gewicht errechnete Größe!) nicht signifikant werden, was ja nach oben Gesagtem lediglich bedeutet, dass er keinen *zusätzlichen*, über Länge und Gewicht hinausgehenden Erklärungswert für die Zielgröße besitzt.
- Das Gesamtmodell ist signifikant, die einzelnen Prädiktoren dagegen nicht: Nach dem oben Gesagten trägt kein Prädiktor über die jeweils anderen hinaus zur Prädiktion der Zielgröße bei, während das Gesamtmodell, i.e. alle Prädiktoren zusammen, durchaus einen hohen Erklärungswert besitzen können.
- Das Gesamtmodell ist nicht signifikant, aber einzelne Prädiktoren: Im letzten Fall ist zu vermuten, dass das verwendete Modell keinen Erklärungswert besitzt und die eventuell gefundenen Signifikanzen möglicherweise durch das multiple Testen bedingte Artefakte darstellen. Andererseits kann auch das Modell ungünstig gewählt sein: Ist z.B. nur *ein* wesentlicher Prädiktor mit ansonsten unkorrelierten Variablen im Modell enthalten, so kann dessen Einfluss im Gesamtmodell überdeckt werden, im Individualtest dagegen erkannt werden.

Die Frage "welche Prädiktoren sind wichtig?" ist im Allgemeinen nicht einfach zu entscheiden und hängt maßgeblich vom inhaltlichen Kontext und in der Regel weniger von den statistischen Eigenheiten bzw. Signifikanzen ab. Die Größe der partiellen Regressionskoeffizienten ist - wegen der unterschiedlichen Streuungen der verschiedenen Parameter - nicht unbedingt maßgeblich, auf manche Parameter kann man Einfluss gewinnen, im Beispiel oben etwa auf den Blutdruck, auf andere dagegen nicht (wie im oben gegebenen Beispiel trivialerweise etwa auf das Alter). Möglicherweise spielen auch Kostenüberlegungen eine Rolle. Prinzipiell sollte die Interpretation des Modells bzw. der Koeffizienten also auch unter medizinisch-inhaltlichen und nicht nur unter statistischen Gesichtspunkten erfolgen.

Zur Vereinfachung eines Modells benutzt man gelegentlich sogenannte Auf- und/oder Abbauverfahren, die in vielen statistischen Programmpaketen (u.a. auch in **BiAS.**) implementiert sind.

- Das *Aufbauverfahren* ("Forward Selection") beginnt mit dem "leeren" Modell und fügt im ersten Schritt die Einflussvariable mit dem kleinsten p-Wert (als die "wichtigste") zum Modell hinzu. Im nächsten Schritt wird diejenige Variable hinzugefügt, die – bei Anwesenheit der bisher aufgenommenen – den kleinsten p-Wert besitzt, mithin also vor dem Hintergrund aller anderen den höchsten *zusätzlichen* Erklärungswert besitzt. Das Verfahren bricht ab, wenn der entsprechende p-Wert eine gewisse Grenze überschreitet, typischerweise γ=0.05 oder 0.10.
- Das *Abbauverfahren* ("Backward Selection") geht umgekehrt vor: Man startet mit *allen* Prädiktoren und entfernt in jedem Schritt jeweils denjenigen mit dem größten partiellen p-Wert, also den vor dem Hinter-

grund aller anderen "unwichtigsten" Prädiktor. Das iterative Verfahren bricht ab, wenn der entsprechende maximale p-Wert kleiner ist als eine gewisse Schranke, typischerweise wieder γ=0.05 oder 0.10.

- Die *Schrittweise Regression* ("Stepwise Regression") ähnelt dem Aufbauverfahren, wobei aber nach Hinzufügen eines Prädiktors möglicherweise bereits vorhandene, im aktuellen Schritt nicht-signifikant gewordene Prädiktoren wieder entfernt werden.

Die genannten Verfahren sind nicht eindeutig und zum Teil umstritten; das Abbauverfahren ist mathematisch am besten abgesichert, so dass man diesem den Vorzug geben sollte (Mantel (1970)). Insbesondere ist auch hier zu bemerken, dass es sich in allen Fällen um statistisch-heuristische Konstruktionen handelt und dass man die Konstruktionshierarchie nach Möglichkeit inhaltlich absichern sollte – oder aber auf die schrittweisen Methoden konsequent verzichtet. Als Ausweg bietet sich das kombinatorische Verfahren "alle möglichen Regressionen" („all subsets“) an, das aber nicht zuletzt wegen des beachtlichen Rechenaufwandes in aller Regel in Programmen nicht verfügbar ist und hier nicht weiter diskutiert wird.

6.11 Logistische Regression

Pragmatisch betrachtet sind die Logistische Regression und die Multiple Lineare Regression (Abschnitt 6.10) nahezu gleichwertig: Beide verwenden ein Modell zur "Vorhersage" einer unabhängigen Zielgröße auf Grundlage einer Anzahl von Prädiktoren, die in beiden Fällen analog interpretiert werden können. Bei Betrachtung aus mathematischer Sicht gibt es aber sehr wesentliche Unterschiede: Die Linearen Regressionsgleichungen können via "Methode der kleinsten Quadrate" (Abschnitt 5.7) explizit gelöst werden, die der Logistischen Regression dagegen nur mit einem iterativen Verfahren, außerdem findet in der Logistischen Regression das Modell der multivariaten Gauß-Verteilung keine Anwendung.

Der prima vista wesentlichste Unterschied der beiden Modelle besteht darin, dass die Lineare Regression eine *quantitative* Zielgröße verwendet, die Logistische Regression dagegen bezieht sich auf eine Transformation einer *nominalen* Indikatorvariablen (zum Beispiel "ja/nein"). Als Einflussgrößen ("Prädiktoren") kommen wie in der Multiplen Linearen Regression sowohl quantitative als auch nominale Größen in Frage, ordinale Prädiktoren können mit Dummy-Variablen codiert werden (vgl. Abschnitt 6.10).

Eine typische Fragestellung der Logistischen Regression richtet sich an das Vorhandensein einer Erkrankung (zum Beispiel der Osteoporose) als Zielgröße in Abhängigkeit von quantitativen und/oder nominalen Prädiktoren als Einflussgrößen (zum Beispiel Alter o.ä.). Vor der Entwicklung der Logistischen Regression in der 1970ern verwendete man häufig die Diskriminanzanalyse (Abschnitt 6.6) in Form eines Multiplen Linearen Modells mit einer 0/1-Variablen (z.B. Erkrankung ja/nein) als Zielgröße, was aber inzwischen von der Logistischen Regression verdrängt wurde:

Anstelle der diskriminanzanalytischen Klassifikation in zwei Gruppen berechnet man mit der Logistischen Regression die geschätzte *Wahrscheinlichkeit*, dass eine Person mit gegebenen Prädiktorwerten zu der einen oder der anderen Gruppe gehört. Eine direkte Modellierung der Indikatorvariablen bzw. der Wahrscheinlichkeiten ist aus weiter unten erläuterten Gründen nicht möglich, so dass man als "Hilfsgröße" die sogenannte *Log-Odds* verwendet. Die Odds eines Ereignisses (siehe auch Abschnitt 6.9) setzt die Wahrscheinlichkeit für einen "Event" (Y=1, Erkrankung o.ä.) in Relation zu "Kein Event" (entsprechend Y=0):

$$\text{Odds}(Y=1) = P(Y=1) / (1-P(Y=1)) = P(Y=1) / P(Y=0)$$

Die "Log-Odds" ist der natürliche Logarithmus der Odds, was gelegentlich auch äquivalent als *Logit-Transformation* bezeichnet wird. Äquivalent zur Berechnung der Log-Odds verwendet man die Logit-Transformation:

$$\text{logit}(P(Y=1)) = \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Die Begründung für die Logit-Transformation ist leicht einzusehen:

Wahrscheinlichkeiten liegen bekanntlich im (beidseitig begrenzten!) Intervall [0,1], wobei für p=0.5 beide Ereignisse gleich wahrscheinlich sind.

Odds liegen im Intervall [0,∞], der neutrale Wert ist 1. Vertauscht man die Rollen der beiden Ereignisse, so erkennt man die Asymmetrie der Odds: Beträgt die Odds für eine Osteoporose für ältere Patientinnen 0.25, so ist die Odds, keine Osteoporose zu haben 1/0.25=4.

Log-Odds bzw. Logits dagegen liegen im Intervall [-∞,+∞], besitzen den neutralen Wert log(0.5/(1-0.5))=log(1)=0 und sind symmetrisch: Bei Vertauschung der beiden Ereignisse ändert sich wegen der Beziehung log(a/b)=–log(b/a) lediglich das Vorzeichen.

Logits sind für die Analyse binärer Zielgrößen bestens geeignet, denn wird die Wahrscheinlichkeit für ein Ereignis größer, so vergrößern sich auch die Logits, wobei beide symmetrisch um einen Neutralwert sind. Insbesondere aber nehmen Logits, wie schon erwähnt, Werte im Intervall von -∞ bis +∞ an, so dass man das allgemeine Logistische Modell für N Einflussgrößen X_i (i=1,2,…,N) mit der Zielgröße logit(p) folgendermaßen definiert:

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_N \cdot X_N$$

Somit können die Schätzwerte b_i der Koeffizienten β_i aus der Logistischen Regression analog zur Multiplen Regression interpretiert werden: Verändert sich die Prädiktorvariable X_i um 1, so bedeutet dies – bei ansonsten gleich bleibenden Werten der übrigen N-1 Einflussvariablen - eine Veränderung der Log-Odds bzw. der Logits um den Betrag b_i.

Mit Hilfe der *Expit-Transformation* bzw. nach einigen Umformungen des allgemeinen Modells ergibt sich die Formel für p=P(Y=1):

$$p = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_N \cdot X_N)]}$$

Dazu eine kurze Vertiefung des oben genannten Osteoporose-Beispiels, das der Darstellung von Dallal (2000/2012) folgt:

Die mit p=P(Y=1)=P(Osteoporose) angepasste Logistische Regressionsgleichung ist

$\text{logit}(p) = b_o + b_1 \cdot X = -4.353 + 0.038 \cdot \text{Alter}$

Da der Koeffizient für X=Alter positiv ist, steigt logit(p) – und damit natürlich auch p – in Abhängigkeit vom Alter an. Über die Exponentialfunktion erhält man die Odds-Ratio p/(1-p),

$p/(1-p) = \exp(-4.353+0.038 \cdot \text{Alter})$

und daraus per Expit-Transformation bzw. nach einigen einfachen Umformungen

$$p = \frac{1}{1 + \exp[-(-4.353 + 0.038 \cdot \text{Alter})]}$$

Dieses Beispiel für nur *einen* Prädiktor (das Alter) ist ein spezieller Fall des allgemeinen Modells, ein Beispiel für N=2 mit einer *Confounder-Variablen* findet sich weiter unten.

Die Koeffizienten b_i des Logistischen Modells können sehr einleuchtend interpretiert werden, wenn man deren transformierte Werte $\exp(b_i)$ untersucht: Vergrößert man eine Einflussgröße X_i um 1, so ist das Odds-Ratio als der Quotient der korrespondierenden Größen $\text{Odds}(Y=1|X_i=x+1)$ und $\text{Odds}(Y=1|X_i=x)$ gerade gleich $\exp(b_i)$.

Im Beispiel der Osteoporose lässt sich dieser Zusammenhang leicht nachvollziehen:

b=0.038 ist der Koeffizient der Einflussgröße X=Alter (Dallal (2000/2012)). Die beiden relevanten Odds für X=x und X=x+1 lauten

$\text{Odds}(\text{Osteoporose}|\text{Alter}=x) = \exp(-4.353+0.038 \cdot x)$
$\text{Odds}(\text{Osteoporose}|\text{Alter}=x+1) = \exp(-4.353+0.038 \cdot (x+1))$

Dividiert man die zweite durch die erste, so erhält man das Odds-Ratio. Nach Umformung via $\exp(-4.353+0.038 \cdot (x+1)) = \exp(-4.353+0.038 \cdot x+0.038) = \exp(-4.353+0.038 \cdot x) \cdot \exp(0.038)$ und Kürzen ergibt sich

$\exp(b) = \exp(0.038) = 1.0387$

Die Odds, dass eine Frau eine Osteporose entwickelt steigt damit um 3.87% pro Jahr. Für eine Altersdifferenz von beispielsweise 10 Jahren (X=x bzw. X=x+10) beträgt das Odds-Ratio $\exp(p)^{10} = \exp(0.038)^{10} = 1.46$, ist also um 46% höher.

Analog zur Multiplen Linearen Regression kann man auch in der Logistischen Regression – mit den gleichen Vorbehalten wie in Abschnitt 6.10! – ein Auf- oder Abbauverfahren durchführen, um zu einem minimalen relevanten Modell zu gelangen.

Berücksichtigt man im Beispiel der Osteoporose zusätzlich zum Beispiel den Prädiktor "Prä/Post-Menopausal" als sog. *Confounder*, so kann man den Einfluss des Alters bezüglich des Menopausalstatus *adjustieren*: Prüft man in dem erweiterten Modell den Einfluss des Alters unter Berücksichtigung des Menopausalstatus, so gibt die Logistische Regression Aufschluss darüber, welchen Einfluss das Alter über die eventuell eingetretene Menopause hinaus auf die Osteoporose besitzt - das Umgekehrte ist natürlich ebenfalls denkbar!

Zur praktischen Durchführung der Logistischen Regression können die Programme **SPSS** und **BiAS.** empfohlen werden, die – einschließlich der Angabe von Konfidenzintervallen! – eine sehr komfortable Behandlung der skizzierten Fragestellungen gestatten; mehr dazu im nächsten Kapitel 7.

6.12 Die Cox-Regression

In Abschnitt 6.7 zur Überlebenszeitanalyse wurde bereits auf die Bedeutung der "Survival-Analyse" bei Therapiestudien und bei prognostischen Fragestellungen hingewiesen. Speziell bei letzteren zeigt sich, dass man eine Prognose in der Regel nicht nur in Abhängigkeit von einer, sondern nur von mehreren Einflussgrößen gleichzeitig beurteilen kann: Typischerweise wird man zusätzlich etwa das Alter, Geschlecht, Schweregrad, TNM-Klassifikation etc. als Kovariablen zur Verbesserung der Prognose heranziehen. Ein Log-Rank-Test, wie in Abschnitt 6.7 skizziert, reicht zur Auswertung solcher komplexen Datenstrukturen nicht aus, so dass man in solchen Fällen eine Analyse mit dem *Cox-Modell* (benannt nach dem zeitgenössischen englischen Statistiker Sir David R. Cox) vornehmen sollte.

Das Cox-Modell ergibt im Beispiel einer Therapiestudie eine Schätzung der Therapieeffekte auf die Überlebenszeit, wobei – wie bereits aus den letzten beiden Abschnitten bekannt – bezüglich der anderen Einflussgrößen adjustiert wird. Damit kann der *Hazard* als das Risiko bzw. die Chance einer Person in Hinblick auf Tod oder ein anderes Zielereignis wie Heilung, Rezidiv oder anderes geschätzt werden.

Für das Cox-Modell sind keine bestimmten Verteilungsvoraussetzungen zu treffen, es ist aber sicherzustellen, dass die Effekte der einbezogenen Kovariablen zeitlich konstant sind.

Die besondere Schwierigkeit bei *time-to-event*-Fragestellungen ist, dass die Zielvariable eine möglicherweise *zensierte* Größe darstellt, das heißt, dass bei manchen Patienten das Zielereignis (zum Beispiel Tod oder Heilung) bereits eingetreten ist, bei anderen dagegen nicht; die Daten der letzteren Patienten bezeichnet man (vgl. dazu Abschnitt 6.7!) als "zensiert".

In den Abschnitten 6.10-11 wurde bereits das Prinzip der Multiplen und der Logistischen Regression vorgestellt. In ähnlicher Weise beschrieb D. R. Cox eine Modellierung nicht unmittelbar der Überlebenszeiten, sondern der sogenannten *Hazard-Funktion* h(t), die von den möglichen Einflussfaktoren bzw. Kovariablen X_1, X_2, …, X_N abhängt:

$$h(t; X_1,X_2,\ldots,X_N) = h_0(t) \cdot \exp\{\beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_N \cdot X_N\}$$

Die Hazardfunktion (die "Ausfallrisikofunktion") gibt das Risiko an, dass ein zum Zeitpunkt t noch lebender Patient unmittelbar danach verstirbt. Aus der Modell-Formel für $h(t;X_i)$ ist ersichtlich, dass die Hazardfunktion einer bestimmten, durch die Kovariablen festgelegten Patientengruppe das Vielfache einer Referenzkurve $h_0(t)$ ("Baseline-Hazard") dargestellt, bei der alle Einflussgrößen $X_i=0$ sind. $h_0(t)$ kommt also eine ähnliche Bedeutung zu wie dem absoluten Glied in einer Multiplen oder Logistischen Regressionsgleichung. Es ist zu beachten, dass die Überlebenszeit umgekehrt proportional zu der Hazard-Funktion ist und umso größer wird, je geringer der Hazard ist.

Die Funktion h(t) kann analog zum Logistischen Modell als eine lineare Gleichung dargestellt werden:

$$\log[\, h(t; X_1,X_2,\ldots,X_N)\,] = \log[h_0(t)] + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_N \cdot X_N$$

Betrachtet man den Quotienten der beiden Hazard-Funktionen zum Beispiel zweier Therapiegruppen A und B (gewissermaßen das "Relative Risiko" A vs. B) und nimmt an, dass dieses über die Zeit betrachtet konstant ist (vgl. oben), so erhält man die Größe

$$HR = h_A(t; X_1,X_2,\ldots,X_N) \,/\, h_B(t; X_1,X_2,\ldots,X_N) = \text{const.}$$

die als *Hazard Ratio* oder als *Relativer Hazard* bezeichnet wird. Zweckmäßigerweise gibt man zusätzlich Konfidenzintervalle an, was prinzipiell sicher nur mit einem geeigneten Rechenprogramm durchführbar ist. Die Signifikanzprüfung erfolgt wie gewohnt über die Konfidenzintervalle oder die entsprechende Prüfgröße bzw. über die zugeordneten p-Werte.

Die zuletzt unterstellte Proportionalität (daher auch die Bezeichnung *proportional hazards model*!) bedeutet zum Beispiel in einer Therapiestudie, dass die Therapieeffekte im Verlaufe der Zeit konstant bleiben, womit auch unterstellt ist, dass die eine der beiden Therapien gleichmäßig besser ist als die andere. Speziell letzteres ist gelegentlich problematisch, da durchaus eine Therapie zu einem frühen – oder auch zu einem späten – Zeitpunkt der anderen überlegen sein kann, während zu anderen Zeitpunkten kein Unterschied vorhanden sein kann. Eine Verletzung der Voraussetzung der Proportionalität kann eine Verwendung anderer Auswerteverfahren empfehlenswert machen.

Die Interpretation der Koeffizienten des Modells ist die gleiche wie im Logistischen Modell (Abschnitt 6.11):

Bei quantitativen Kovariablen erhöht sich der Hazard um den Betrag $\exp(\beta_i)$, wenn sich die Einflussgröße X_i um 1 vergrößert. Bei nominalen Einflussgrößen X_i (zum Beispiel: Metastasen ja/nein bzw. 1/0) bedeutet $\exp(\beta_i)$ die Erhöhung des Hazards bei Vorhandensein von Metastasen.

Bei ordinalen Skalen oder bei mehreren Kategorien behilft man sich mit Dummy-Variablen: Liegt zum Beispiel eine Kovariable "Stadium" mit den Stufen 1, 2, 3 und 4 vor, so definiert man drei Dummy-Variablen S2, S3 und S4 und benutzt das Stadium 1 als Referenzkategorie: $\exp(\beta_{S3})$ etwa bedeutet dann die Erhöhung des Hazard für Patienten im Stadium 3, dies relativ zur Referenzkategorie betrachtet.

Klein und Moeschberger (1997) stellen als Beispiel die Stilldauer der Mutter in Abhängigkeit von einer Reihe von Einflussgrößen dar: Diese sind Wohlstand, Rauchen, Alter, Dauer der Ausbildung und Ethnizität. Als Ergebnis der Cox-Regression kann man zum Beispiel das Hazard Ratio ermitteln, dass – bei ansonsten gleichen Einflussfaktoren - das Risiko des Abstillens für eine Raucherin im Vergleich zu einer Nichtraucherin um den Faktor 1.28 höher ist. Wegen der Nähe zur Logistischen Regression wird auf eine ausführliche Interpretation verzichtet.

Zum Cox-Modell kann man wie in der Multiplen und Logistischen Regression einen schrittweisen Auf- oder Abbau des Modells durchführen, was hier ebenfalls nicht mehr eingehend erläutert werden muss.

Zum praktischen Rechnen empfehlen sich die Programmpakete **SPSS** und **BiAS**.. Ausführlicheres findet sich bei Klein und Moeschberger (1997).

6.13 Der Propensity-Score

In den ersten Kapiteln dieses Skriptums wurde als Ziel einer kontrollierten klinischen Studie (CCT/RCT, *controlled/randomized clinical trial*) der Vergleich von zwei oder mehreren strukturgleichen Behandlungsgruppen diskutiert, wobei "Strukturgleichheit" die Vergleichbarkeit der Studienarme bezüglich aller relevanten Ausgangs- und Behandlungsbedingungen bedeutet. Die Methode der "Randomisierung" (Abschnitt 2.4) garantiert die gewünschte Strukturgleichheit und gewährleistet die Anwendbarkeit der bisher in diesem Skriptum beschriebenen Methoden.

Der komplementäre Studientyp einer epidemiologischen Kohortenstudie vergleicht dagegen *nicht-randomisierte* Studiengruppen, wie dies zum Beispiel bei retrospektiven "Beobachtungsstudien" der Fall ist. Bei solchen Studien sind die Behandlungen bzw. ärztlichen Maßnahmen nicht "per Zufall" zugeteilt, sondern werden in Abhängigkeit von unterschiedlichen Kriterien (Einflussvariablen oder "Kovariablen", auch: "*Confounder*") nach ärztlicher Entscheidung oder sogar unter Beteiligung des Patienten vorgenommen. Die genannten Ko- oder Confoundervariablen werden also im Allgemeinen ihrerseits einen Einfluss auf den Behandlungserfolg besitzen: Damit sind die Behandlungsgruppen nicht mehr strukturgleich und ein unmittelbarer Vergleich der Behandlungsgruppen ist ebensowenig möglich wie eine "klassische" Auswertung solcher Studien.

Als erste Lösungen kommen prinzipiell eine "Stratifizierung" oder eine "Kovarianzanalyse" in Frage. Im ersten Fall betrachtet man bezüglich der Kovariablen vergleichbare Untergruppen oder verwendet die "Matched-Pairs-Technik", bei der jedes "Paar" jeweils aus zwei Patienten mit unterschiedlichen Behandlungen, aber mit möglichst ähnlichen Kovariablen besteht. Im zweiten Fall werden die relevanten Kovariablen in einem speziellen Modell der Varianzanalyse erfasst und somit die bezüglich der Kovariablen bereinigten ("adjustierten") Behandlungseffekte verglichen.

Die "Kovarianzanalyse" dient wie der in Kapitel 5 angesprochene Zweistichproben-t-Test und die Varianzanalyse (letztere als Verallgemeinerung des t-Tests für drei oder mehrere Behandlungsgruppen) dem Vergleich von zwei oder mehr Gruppen, wobei man dabei auch fragliche Störgrößen im Sinne von Kovariablen bei der Analyse berücksichtigen kann.

Diese "*Adjustierung*" (Analoges gilt für die Subgruppenanalyse) ist sicher umso besser, je mehr Kovariablen erfasst werden, womit dieser Ansatz rasch an praktische Grenzen der Durchführbarkeit stößt. Rosenbaum und Rubin (vgl. D'Agostino (1998)) fanden eine auch unabhängig von der vorliegenden Fallzahl immer praktikable Alternative, die im Allgemeinen als "*Propensitiy-Score*" oder als "*Zuteilungsscore*" bezeichnet wird:

Dieser Score ist eine Funktion *aller* relevanten Kovariablen, die zur Zuteilung der einen oder der anderen Behandlung geführt haben können. Ein naheliegender Score ist die Wahrscheinlichkeit für die Zuteilung einer bestimmten der beiden Behandlungen als Funktion der Kovariablen, wobei diese Wahrscheinlichkeit recht einfach per Diskriminanzanalyse (vgl. 6.6)

oder auch per Logistische Regression (Abschnitt 6.11, mit der Zielgröße "Behandlung") berechnet werden kann. Es zeigt sich, dass bei der oben angesprochenen Stratifikation bzw. bei der Kovarianzanalyse mit Hilfe des Propensity-Scores ein optimaler Ausgleich erzielt wird, wenn die betrachtete Funktion *alle* für die Zuteilung der Behandlungen relevanten Kovariablen enthält. Ist dies der Fall, so ist die Analyse per Propensity-Score als *einzige* (!) Kovariable genauso valide wie nach einer Randomisierung.

Vielfach wird die Überlegenheit einer "Propensity-Analyse" gegenüber von klassischen randomisierten Studien behauptet, da Kohortenstudien ein eher unverfälschtes Bild der medizinischen Realität bieten und die praktische Anwendung von Arzneimittel besser abbilden. Als Kritikpunkt ist zu bedenken, dass möglicherweise relevante *Confounder-Variablen* übersehen werden können, nicht bekannt sind oder wegen des retrospektiven Charakters einer Studie nicht mehr ermittelbar sind und somit – speziell im Vergleich mit CCTs! – zu einer Verfälschung der Resultate beitragen.

Es wird vermutet, dass eine Katarakt-Operation als Risikofaktor für die Entwicklung einer feuchten altersabhängigen Makuladegeneration (sog. "fAMD", mit dem Risiko der Erblindung) aufzufassen ist. Die Behandlung des grauen Stars kann aber aus sachlichen und ethischen Gründen nicht randomisiert vorgenommen werden, sondern die Entscheidung wird vom Arzt in Abhängigkeit u.a. von den "Kovariablen" Alter, Visus etc. getroffen. Nach dem oben Diskutierten berechnet man also mit Hilfe der Diskriminanzanalyse oder der Logistischen Regression zunächst den Propensity-Score in Abhängigkeit von Alter, Visus etc. als die Wahrscheinlichkeit, dass ein Patient operiert wird. Im zweiten Schritt erfolgt eine erneute Logistische Regression mit "fAMD" als Zielgröße bzw. "Behandlung ja/nein" und Propensity-Score als Einflussgrößen; der Propensity-Score ist dabei die ***einzige*** Kovariable.

Gemäß Kapitel 6.11 zur Logistischen Regression ergibt sich eine Beurteilung des Einflusses der Katarakt-Operation, der über den „Einfluss“ des Propensity-Scores (beziehungsweise über den Einfluss der beteiligten Kovariablen/Confounder) hinaus die Entwicklung einer feuchten Makuladegeneration begünstigt: In diesem Sinne werden die Ergebnisse bezüglich der Kovariablen „adjustiert“ bzw. von deren Einfluss bereinigt.

Zur weiteren Interpretation stelle man sich vor, dass prinzipiell alle älteren Patienten mit schlechtem Visus operiert werden würden, alle anderen dagegen nicht: Damit wäre die fAMD alleine durch den Propensity-Score erklärbar, die OP würde – im Sinne von Kapitel 6.11 – keinen weiteren Beitrag bzw. Erklärungswert für die Entstehung einer AMD leisten können.

Das Prinzip des Propensity-Scores lässt sich naheliegenderweise auch bei einer Analyse per Multipler Regression (Abschnitt 6.11) oder auch in der Überlebenszeitanalyse – zum Beispiel im Cox-Modell aus Abschnitt 6.12 – anwenden, was aber hier nicht weiter ausgeführt wird. Interessante Beispiele finden sich in der Arbeit von Schneider (2001), die auf mehrere Beobachtungsstudien hinweist, in denen die Nützlichkeit des Propensity-Zuteilungsscores unter Beweis gestellt wurde.

Die Analyse von epidemiologischen Kohortenstudien mit Hilfe des dargestellten Propensity-Scores wird inzwischen – jedenfalls unter bestimmten Bedingungen – als gültiger Nachweis der Wirksamkeit und der Arzneimittelsicherheit anerkannt. In den "Empfehlungen zur Planung und Durchführung von Anwendungsbeobachtungen" (1998) des BfArM finden sich Einzelheiten dazu.

6.14 Sequentielle und Adaptive Designs

In vielen klinischen Studien möchte man schon vor Erreichen der geplanten Fallzahl eine *Zwischenauswertung* ("Interims-Analyse") vornehmen, da sich eine Studie in Abhängigkeit von dem Studiengegenstand (man denke nur als Beispiel an Überlebenszeitstudien) unter Umständen über einen sehr langen Zeitraum erstreckt. Zwischenauswertungen können dadurch motiviert sein, dass man vielleicht bereits zu einem früheren Zeitpunkt als geplant die gewünschten Effekte zeigen kann, damit die Studie abbrechen kann und somit bereits viel früher zukünftige Patienten von den Studienerkenntnissen profitieren. Eine weitere Motivation für einen früher als geplanten Studienabbruch kann auch die Analyse von Nebenwirkungen ("*adverse effect*", "*serious adverse effect*") sein. Sequentielle und adaptive Designs bzw. Tests tragen diesem Konzept Rechnung.

Eine sehr simple, allerdings auch sehr konservative Methode ist bereits aus Abschnitt 5.8 durch das Multiple Testen bekannt: Möchte man zum Beispiel vor der Endauswertung N-1=4 Zwischenauswertungen vornehmen, so kann man jede der insgesamt N Auswertungen an der Bonferroni-adjustierten Signifikanzschwelle $\alpha^*=\alpha/N$ vornehmen, um für die gesamte Studie ein gewünschtes α von z.B. 0.05 zu garantieren.

Wie aus der Bonferroni-Korrektur ersichtlich ist, wird es bei steigender Anzahl der Zwischenauswertungen natürlich immer schwieriger, an der korrigierten Schwelle α^* eine Ablehnung der Nullhypothese zu erzielen, oder, umgekehrt, erfordert dies bei gegebener Wahrscheinlichkeit für den Fehler 2. Art bzw. bei gegebener Power $1-\beta$ und einer nachzuweisenden minimalen Differenz $\delta=\mu_1-\mu_2$ (vgl. Abschnitte 4.5 bzw. 5.10) eine unter Umständen erheblich größere Fallzahl n. Damit kann die gute Absicht der Interims-Analyse im Grenzfall sogar ein positives Resultat einer Studie verhindern. Inzwischen gibt es aber eine ganze Reihe modernerer Verfahren, die diesem Umstand Rechnung tragen:

Pocock (1977) stellte ein erstes sequentielles Verfahren vor und konnte damit eine weniger konservative α-Korrektur als Bonferroni begründen; ein Nachteil dieser Methode ist, dass das gleiche α^* für alle Auswertungen verwendet wird. O'Brien und Fleming (1979) gelang eine Flexibilisierung dieses Vorgehens, wie aus dem folgenden Auszug einer Tabelle für eine Gesamt-Irrtumswahrscheinlichkeit α von 0.05 ersichtlich ist:

N=2		N=3			N=4			
alpha-1	alpha-2	alpha-1	alpha-2	alpha-3	alpha-1	alpha-2	alpha-3	alpha-4
0.00500	0.04806	0.00250	0.00296	0.04831	0.00167	0.00194	0.00233	0.04838
0.01000	0.04519	0.00500	0.00612	0.04588	0.00333	0.00400	0.00488	0.04614
0.01500	0.04177	0.00750	0.00936	0.04292	0.00500	0.00612	0.00753	0.04342
0.02000	0.03788	0.01000	0.01268	0.03949	0.00667	0.00827	0.01026	0.04025
0.02500	0.03355	0.01250	0.01606	0.03558	0.00833	0.01047	0.01306	0.03660

Tabelle 20: Interims-Analyse nach Fleming und O'Brien für α=0.05

In der Tabelle bedeutet N wieder die Anzahl der Auswertungen, angegeben sind jeweils die nominalen Signifikanzniveaus α_i=alpha-i (i=1,..,N) der N-1 Interimsanalysen und der Endauswertung.

Führt man gemäß Tabelle 20 N-1=2 Interims-Analysen durch, so kann man die erste bei α_1=0.00750, die zweite bei α_2=0.00936 und die Endauswertung bei α_3=0.04292 durchführen: Bei unerwartet großen Effekten (zum Beispiel der Differenz zwischen zwei Gruppen) kann man gegebenenfalls schon zu einem frühen Zeitpunkt die Studie mit positivem Ergebnis abbrechen, falls dies nicht der Fall sein sollte, kann man aber – und dies im Unterschied zu Pocock oder Bonferroni! – die Endauswertung mit einem Wert $\alpha_N=\alpha_3$ durchführen, der nahezu der Irrtumswahrscheinlichkeit $\alpha_{gesamt}=\alpha$ entspricht. Andere Kombinationen aus der zweiten Spalte der Tabelle sind natürlich, falls adäquater, ebenfalls denkbar.

In sequentiellen Designs testet man die Abfolge der Nullhypothesen sequentiell unter Einbeziehung der gerade vorliegenden Daten, in adaptiven Designs in jeder Phase separat. Im Verlauf der Zwischenauswertungen kombiniert man die in jedem Schritt erhaltenen p-Werte und gelangt damit zu einer Gesamtaussage im Sinne eines globalen p-Wertes. Damit entfällt die Festlegung auf eine konkrete Irrtumswahrscheinlichkeit für jede Zwischenauswertung, und, darüber hinaus, kann man nach jeder Interims-Analyse das Studiendesign ändern und insbesondere auch die Fallzahlen für die nachfolgenden Studienabschnitte neu berechnen. Dieses Studiendesign, das auf Bauer und Köhne (1994) zurückgeht, wird im Folgenden anhand eines Beispiels für zwei Auswertungsstufen erläutert.

Alle bisher erwähnten Designs sind in dem Programmpaket **BiAS.** verfügbar: Die Methoden von Pocock und Fleming-O'Brien sind in Form von Tabellen verfügbar, das aufwendigere Design von Bauer und Köhne kann über einen Eingabedialog für jeden Studienabschnitt neu berechnet werden. Das nachfolgende Beispiel zum Bauer-Köhne-Design wurde mit Hilfe von **BiAS.** durchgerechnet:

Für die - hier einzige - Zwischenauswertung gibt man eine Irrtumswahrscheinlichkeit α_0 vor, für die man die Studie in jedem Fall wegen Nutzlosigkeit der neuen Therapie abbrechen will ("*Stop-for-futility*"), womit die Nullhypothese beibehalten würde. Auf dieser Grundlage wird für die erste Phase der Studie die Schwelle α_1 berechnet, an der die Studie bereits in dieser Phase zu Gunsten von H_A beendet werden kann. Falls der p-Wert p_1 *zwischen* α_0 und α_1 liegt, folgt die zweite Phase der Studie, wozu man die erforderliche lokale Irrtumswahrscheinlichkeit α_2 berechnet, die das vorgegebene globale Signifikanzniveau α der Studie garantiert.

Dazu ein Beispiel: α wird mit 0.05 festgelegt, α_0 als absolute Nutzlosigkeitsgrenze mit 0.30. Für die erste Studienphase errechnet sich nach Bauer und Köhne ein Signifikanzniveau von α_1=0.029938, das mit einem p-Wert von p_1=0.12 zwar nicht erreicht wird, aber doch kleiner ist als α_0=0.30, womit die Studie fortgesetzt wird. (Nach diesem Studienabschnitt kann das Design der Studie geändert werden, insbesondere kann auch eine Neuberechnung der Fallzahlen vorgenommen werden.) Es erfolgt eine Neuberechnung des Signifikanzniveaus α_1 für den zweiten Studienabschnitt mit α_1=0.072542. Im Beispiel möge der entsprechende p-Wert p_2=0.068 betragen, woraus sich ein Gesamt-p-Wert von p=0.04797 ergibt, der kleiner ist als α=0.05 und damit zur Ablehnung der globalen Nullhypothese führt.

In der neueren Literatur finden sich weitere Methoden, die in diesem Skriptum jedoch nicht weiter behandelt werden können.

Kapitel 7:
Statistische Programmpakete

Zur Datensammlung und für erste deskriptive Auswertungen gibt es eine ganze Reihe von *Datenbank- und Tabellenkalkulationsprogrammen* wie zum Beispiel die Klassiker **Excel**, **Access**, **dBase** oder auch **OpenOffice**. Solche und andere, hier nicht genannte Programme sind für die Datenverwaltung gut geeignet und ausgereift, so dass eine Auswahl eher eine Frage des persönlichen Geschmacks und weniger eine sachliche Frage ist. Ein nicht zu unterschätzendes Auswahlkriterium sind natürlich auch Freunde oder Kollegen, die bereits Erfahrung mit einem Programm besitzen und vielleicht Tipps und Unterstützung bei der oft nicht einfachen Einarbeitung anbieten können. Bitte beachten Sie aber, dass die oben erwähnten Programme keine oder nur rudimentäre Funktionen für eine sachgerechte teststatistische Auswertung besitzen.

Ein ganz wesentliches Kriterium bei der Auswahl eines Datenbank- oder Tabellenkalkulationsprogramms ist die Möglichkeit des Datenexports, um auch anderen Programmen - zum Beispiel Graphikprogrammen und Statistik-Paketen - die erarbeitete Datenbank zur Verfügung stellen zu können. Es ist günstig, wenn man wenigstens im *dBaseIII*- oder in einem *ASCII-Format* (zum Beispiel in den Windows-Formaten *CSV* oder *TXT*) Daten exportieren kann. Das Umgekehrte gilt natürlich auch für Graphik- und für Statistik-Programme: Diese sollten wenigstens die genannten Formate importieren, also lesen können. Optimal ist die Möglichkeit eines Datenaustauschs über die Zwischenablage (*Clipboard*).

Zur graphischen Darstellung der Daten haben sich Programme wie **Excel**, **PowerPoint**, **GraphpadPrism** oder auch **HarvardGraphics** bewährt. Unabhängig davon sollte ein Programm, das man vielleicht erwerben möchte, die bereits erwähnte Möglichkeit des Imports von Fremddateien wenigstens in einem der genannten Formate besitzen. Bei Graphik-Programmen ist nach Ansicht des Autors auch auf die Möglichkeit des Exports bzw. Schreibens von *BMP-Graphikdateien* (Bitmap) und/oder *GIF*- und *JPG-Graphikdateien* zu achten; dadurch ist sowohl eine Dokumentation als auch ein Import in Textprogramme wie **Word** gegeben.

Statistik-Programme gibt es inzwischen zahlreiche, allerdings auch zahlreiche wenig empfehlenswerte. Sehr empfehlenswert, allerdings aber nicht "billig" sind die Klassiker **SPSS**, **SAS** oder **BMDP**. Diese Programme sind sehr umfangreich und gehen, jedenfalls im teststatistischen Bereich, zum Teil weit über das hinaus, was ein "durchschnittlicher" Anwender bei seiner Arbeit benötigt - in Hinblick auf die Anwendung in der Medizin fehlen aber leider auch häufig spezielle Methoden, die nach Ansicht des Autors unbedingt Bestandteil eines statistischen bzw. biometrischen Programmpaketes sein sollten: Das sind z. B. Exakte Tests in der Nicht-

Parametrik, "Post-hoc-Tests" für z.B. Kruskal-Wallis-Test, Fallzahl- und Powerberechnungen und vieles andere mehr. Andererseits sollten ansprechende Benutzeroberflächen und/oder **Excel**-Kompatibilität nicht von einer kritischen Auseinandersetzung mit dem Umfang, der Adäquatheit und der Qualität der statistischen bzw. biometrischen Inhalte abhalten. Weitere, ebenfalls nicht unbedingt preisgünstige Programme wie zum Beispiel **TESTIMATE** (IDV München) sind sicher ebenfalls zu empfehlen, wurden aber mangels Verfügbarkeit nicht vom Autor getestet. Im Internet finden Sie zahlreiche weitere interessante Hinweise zu vielen guten Programmen wie **R**, **MiniTab**, **SYSTAT**, **BioStat**, **MedCalc** und viele andere mehr.

In der preisgünstigeren Klasse ist Vorsicht geboten, denn hier gibt es erhebliche Qualitätsunterschiede. In jedem Fall sollte man in dieser Preisklasse darauf achten, dass die oben genannten Kriterien des Datenimports und -exports gewährleistet sind, denn speziell bei kleineren Paketen (aber auch bei "großen"!) erlebt man oft die Überraschung, dass ein bisher vielleicht gerne benutztes Programm doch einiges vermissen lässt und man deshalb auf ein anderes ausweichen muss: Eine Umsetzung von Dateiformaten ist jedoch grundsätzlich problematisch und mitunter auch fehleranfällig. Das Programm **DBMS/Copy** leistet hervorragende Dienste beim Konvertieren.

Als weiteres Auswahlkriterium sollte man unbedingt das Vorhandensein von Unterlagen zur Programm-Validierung zur Bedingung machen: Alle im Programm angebotenen Verfahren sollten wenigstens in Form von Beispielen aus der Literatur und/oder via Referenzprogramme überprüfbar sein. So ist dem Autor ein Excel-Add-On bekannt, das bereits bei der Berechnung der Standardabweichung eigene, nicht nachvollziehbare Wege geht. Ein weiteres Programm versagte im Test bereits bei der Berechnung des χ^2-Vierfelder-Tests. Auf alle Fälle wird man die Ergebnisse eines neu erworbenen Programms mit vielleicht vorhandenen eigenen Ergebnissen oder Ergebnissen aus der Literatur vergleichen, um eventuelle Unstimmigkeiten oder vielleicht sogar fehlerhafte Berechnungen zu entdecken.

Als letzte, nicht unbedingt weitreichende Bedingung sollte ein statistisches Programmpaket nicht nur sogenannte asymptotische Tests, sondern auch exakte Tests in der Nicht-parametrischen Statistik anbieten (zur Erinnerung: Wilcoxon-Mann-Whitney-Test!). Asymptotische Testvarianten gelten nur für größere Fallzahlen und können bei kleinen Stichprobenumfängen zu nicht unerheblichen Fehlern führen. Auf das Erfordernis eines Handbuches muss sicher nicht eigens hingewiesen werden.

Das Programmpaket **BiAS.** ("**Bi**ometrische **A**nalyse von **S**tichproben"), vom Autor dieses Skriptums entwickelt, ist ein Programm in der unteren Preisklasse und erfüllt alle oben formulierten Forderungen. **BiAS.** wurde für Mediziner, Biologen und Psychologen aus der Sicht eines Biometrikers entworfen und beinhaltet eine umfassende, den Bedürfnissen seiner Benutzer angemessene Auswahl von Methoden der deskriptiven und der konfirmatorischen Statistik. **BiAS.** ist ab Windows 95 bis Windows 10 lauffähig,

kann aber auch mit Hilfe der Emulatoren *Wine*, *Parallels* oder *VirtualBox* (oder auch *ohne* Emulator mit WineBottler!) unter den Betriebssystemen *Mac OS X*, *Linux*, *Ubuntu, Solaris* und anderen Systemen eingesetzt werden. **BiAS.** besitzt ein integriertes Hypertext-Handbuch und ein äquivalentes eBook mit etwa 250 Seiten, beide mit einem umfangreichen Kapitel zur Validierung des Programms. Im Programm ist zusätzlich ein technisches und ein davon getrenntes biometrisches Hilfesystem mit Beispielen zu allen Verfahren verfügbar. Eine Testversion inclusive Hypertext-Handbuch ist unter https://www.bias-online.de per Download erhältlich.

BiAS. enthält neben den in diesem Buch vorgestellten "Standardverfahren" zahlreiche weitere Module, die in anderen, auch größeren Programmpaketen nicht oder nicht ohne Weiteres verfügbar sind: Dies sind ein Datenbankmodul, Fallzahl- und Power-Berechnungen, diverse Randomisierung, multivariate Verfahren für zwei Gruppen, Multiple und Logistische Regression (auch mit Abbau-Verfahren), Konfigurationsfrequenzanalyse (KFA), Diagnostische Tests, parametrische und nicht-parametrische, auch nominale Cross-Over-Analyse, Survival-Analyse, Cluster- und Diskriminanzanalyse (letztere ebenfalls mit einem Abbau-Verfahren), Zeitreihenanalyse, Toleranzbereiche, Bioäquivalenzprüfung, Faktorenanalyse, ROC-Kurven und anderes mehr. Beispiele für **BiAS.**' Graphiken finden sich in fast allen Kapiteln dieses Skriptums.

BiAS. erlaubt den Import und Export von **Excel**-, **dBase**-, **SPSS**- und ASCII-Dateiformaten (*.XLS, *.DBF, *.SAV, *.TXT, *.CSV und *.SDF) und kann auf diesem Wege Dateien mit z.B. **Excel** austauschen. Graphiken können als Hardcopy, per Zwischenablage oder als BMP-, GIF-, JPG- und PCX-Dateien ausgegeben und in geeigneter Weise für das Einlesen in zum Beispiel Textprogramme wie **Word** vorbereitet werden. Die Graphik-Dateien können vom Programm aus auch beispielsweise MS-**Paint** oder **PowerPoint** zur Verfügung gestellt werden.

Statistik-Programme werden vielfach verwendet wie Textprogramme. Unabhängig davon, für welches Statistik-Paket man sich entscheidet, sollte man sich verinnerlichen, dass die Verfügbarkeit über ein statistisches Programmpaket noch nicht die Verfügbarkeit über dessen wissenschaftliche Inhalte bedeutet. Der vorliegende Text hat sicher deutlich gemacht, dass Statistik bzw. Biometrie nicht nur einfach Rechnen bedeutet, sondern dass dieses Fachgebiet sehr viel mehr umfasst als die Benutzeroberfläche eines Programms ad hoc erkennen lässt. Sinnloses Rechnen unter falschen Voraussetzungen hat sicher keine vertretbare Prognose, so dass der Autor dringend empfiehlt, vor der ersten Anwendung einer vielleicht bisher unbekannten statistischen Methode ein einschlägiges Lehrbuch zu konsultieren; so wurde in diesem Text bereits mehrfach auf das Lehrbuch von Lothar Sachs (Springer-Verlag 2004/2018) zur weiteren Lektüre oder auch einfach zum Blättern und Informieren hingewiesen: Das Buch beinhaltet viele nützliche Informationen und zahlreiche konkrete Beispiele einschließlich der statistischen Interpretation der Berechnungen.

Anhang:
Mathematische Grundlagen

In vielen Lehr- und Studienplänen medizinnaher Ausbildungsrichtungen ist eine eigene Mathematik-Vorlesung vorgesehen, in anderen nicht: Wenn man aber über Biometrie oder Medizinische Statistik spricht, muss man zwangsläufig - denn immerhin ist die Statistik ein Teilgebiet der Mathematik - auch einiges Wissen über mathematische Inhalte voraussetzen. Im Folgenden kann und soll aber nicht der Versuch unternommen werden, auf zehn Seiten alle erforderlichen mathematischen Grundlagen zu vermitteln, die man bis zum Abitur in der Schule lernt, sondern, für die einen lediglich zur Erinnerung an früher erlerntes Wissen, für die anderen eher als Einstieg in die mathematische Materie, finden sich hier in loser Folge einige wichtige Sachverhalte, denen man in der Statistik immer wieder begegnet. Die Darstellung, keineswegs "vollständig" und vielfach nur in Form von Beispielen vorgenommen, dient somit eher zum Nachschlagen, weniger aber zum systematischen Erlernen der behandelten Gegenstände. Zur weiteren, vertiefenden Lektüre empfiehlt es sich deshalb, ein vielleicht noch vorhandenes Schulbuch oder, wenn man Kompaktes liebt, zum Beispiel das "Abiturwissen Mathematik" von Harald Scheid (2012, Klett-Verlag) oder Ähnliches heranzuziehen. Falls erforderlich, wird jedoch in allen Kapiteln auch etwas ausführlicher auf speziellere mathematische Probleme eingegangen.

Da die Inhalte dieses Kapitels zwar nicht recht zu einer Darstellung der Medizinischen Biometrie passen, aber doch vielfach eine Grundlage zum Verständnis darstellen, wurden sie behelfsweise in diesen Anhang verbannt. Unabhängig davon wird versucht, zu allen mathematischen Inhalten einen unmittelbaren Bezug zur biologischen Anwendung herzustellen.

A.1 Zahlen

In der Mathematik unterscheidet man verschiedene Arten von Zahlen. Die komplexen Zahlen sind dabei - von etwa Quaternionen etc. abgesehen! - die allgemeinste Form, die sich wiederum aus den reellen und den imaginären Zahlen zusammensetzen. Während sich die imaginären Zahlen nicht weiter unterteilen lassen, bestehen die reellen Zahlen aus den rationalen und den irrationalen Zahlen, letztere wiederum beinhalten als Spezialfall die transzendenten Zahlen. Rationale Zahlen setzen sich aus Brüchen und den ganzen Zahlen zusammen, wobei die ganzen Zahlen aus den positiven Zahlen, der Null und den negativen Zahlen bestehen. Diese unterschiedlichen Zahlenmengen können wie folgt definiert werden:

Unter den *natürlichen Zahlen* versteht man die Zahlen 1,2,3,... bis "unendlich". Das bedeutet, dass es zu jeder natürlichen Zahl n einen Nachfolger n+1 gibt, der um 1 größer ist als n, und dies unabhängig davon, wie groß die Zahl n gewählt wird. Man gelangt somit nie an eine endgültig größte natürliche Zahl, wie durch das mathematische Symbol "∞" angedeutet wird. Als Symbol für die natürlichen Zahlen verwendet man in aller Regel das Symbol $\boldsymbol{N}$. Die um die Zahl Null erweiterte Menge wird mit $\boldsymbol{N_0}$ bezeichnet.

Die *ganzen Zahlen* umfassen die natürlichen Zahlen, die Zahl Null und alle mit negativen Vorzeichen versehenen natürlichen Zahlen, bestehen also aus den Zahlen ...,-3,-2,-1,0,1,2,3,... von $-\infty$ bis $+\infty$. Diese Zahlen werden oft mit dem Symbol $\boldsymbol{Z}$ bezeichnet.

Den Quotienten zweier beliebiger ganzer Zahlen bezeichnet man als *Bruch*. Als Ausnahme ist die Division durch Null nicht definiert.

Die Zahlen, die sich aus den ganzen Zahlen und den Brüchen zusammensetzen, bezeichnet man als *rationale Zahlen*. Diese werden häufig mit dem Symbol $\boldsymbol{Q}$ bezeichnet.

Zahlen, die nicht durch einen Bruch darzustellen sind, werden *irrationale Zahlen* genannt. Irrationale Zahlen sind z.B. $\sqrt{2}$, die Zahl π oder die Eulersche Zahl e als Basis der natürlichen Logarithmen. Irrationale Zahlen besitzen grundsätzlich unendlich viele Dezimalstellen, wenn auch nicht jede Zahl mit unendlich vielen Dezimalstellen eine irrationale Zahl ist: Man denke dabei nur an die Zahl 1/3, die zwar unendlich viele Stellen besitzt, aber als Bruch eine rationale Zahl ist.

Die *reellen Zahlen* $\boldsymbol{R}$ setzen sich aus den rationalen und den irrationalen Zahlen zusammen.

Über *transzendente Zahlen* und über *imaginäre Zahlen* muss im Rahmen dieses Skriptums nicht gesprochen werden, da diese in der hier behandelten Statistik keine explizite Rolle spielen.

Die Medizinische Biometrie beschäftigt sich mit Messwerten und mit Beobachtungen, die man im biologischen und medizinischen Bereich erhält. Zum Verständnis sind im Wesentlichen die Charakteristika der natürlichen und der reellen Zahlen ausreichend.

Natürliche Zahlen findet man speziell beim "Abzählen" von Ereignissen vor, reelle Zahlen spielen prinzipiell bei allen Messwerten im Labor eine Rolle. Die Statistik spricht dabei in der Regel von diskreten und von stetigen Skalen, deren Definition - ganz ähnlich zu der eben vorgenommenen - in Abschnitt 1.1 ausgeführt wird. Im gleichen Abschnitt findet sich auch eine Definition von nicht-quantitativen Daten, die hier noch nicht angesprochen wurden. Dies sind *Ordinale* und *Nominale Daten*, die lediglich Rangordnungsinformation bzw. eine kategoriale Klassifikation beinhalten: Im Gegensatz zu quantitativen Daten kann man mit solchen Daten keine Rechenoperationen wie Addition oder Multiplikation durchführen.

A.2 Mengenlehre

Speziell in der Wahrscheinlichkeitsrechnung werden häufig mengentheoretische Begriffe und Symbole verwendet, so dass sich auch hier ein kurzer Exkurs lohnt.

Im letzten Abschnitt war bereits von Mengen die Rede: Zum Beispiel ist die Menge der natürlichen Zahlen ***N*** definiert durch die Zahlen 1,2,3,... . Damit kann man auch sagen, dass n=17 ein Element der natürlichen Zahlen ist, symbolisch: $n \in \boldsymbol{N}$.

Die Menge ***N*** ist eine *Teilmenge* der Menge der ganzen Zahlen ***Z***, symbolisch: $\boldsymbol{N} \subset \boldsymbol{Z}$. Das Umgekehrte ist nicht der Fall, also: $\boldsymbol{Z} \not\subset \boldsymbol{N}$.

Mengen schreibt man auch häufig in geschweifte Klammern. Zum Beispiel ist {1,2,3,4,5,6} die Menge der Zahlen, die man beim Würfeln erhalten kann. Die Reihenfolge der Elemente einer Menge spielt weder in der Schreibweise noch inhaltlich eine Rolle; z.B. ist {a,b,c}={b,c,a}. Als Spezialfall gibt es auch eine leere Menge. Dies ist eine Menge, die kein Element enthält, eben *leer* ist. Als Schreibweise für die *leere Menge* verwendet man ϕ oder einfach { }.

Mengen kann man auch verknüpfen. Die *Vereinigung* bedeutet, dass die resultierende Menge beide zu vereinigende Mengen enthält. Symbolisch schreibt man z.B. $\boldsymbol{C}=\boldsymbol{A} \cup \boldsymbol{B}$. Enthält z.B. ***A*** die Zahlen 1,2,3 und ***B*** die Zahlen 2,7,8, so ist $\boldsymbol{C}=\{1,2,3,7,8\}$.

Möchte man feststellen, welche Elemente zweier Mengen ***A*** und ***B*** in beiden Mengen - also sowohl in ***A*** als auch in ***B*** - enthalten sind, so bildet man die *Schnittmenge*, symbolisch: $\boldsymbol{C}=\boldsymbol{A} \cap \boldsymbol{B}$. Enthält z.B. wieder die Menge ***A*** die Zahlen 1,2,3 und die Menge ***B*** die Zahlen 2,7,8, so ist die Schnittmenge $\boldsymbol{C}=\boldsymbol{A} \cap \boldsymbol{B}=\{2\}$.

Ist die Schnittmenge zweier Mengen leer, so enthalten die beiden Mengen keine gemeinsamen Elemente und werden deshalb auch *elementfremd* oder *disjunkt* genannt.

Vielfach macht man sich Mengen mit Hilfe sogenannter *Venn-Diagramme* anschaulich. Venn-Diagramme umschließen die Elemente einer Menge durch eine geschlossene Linie, wie die Darstellung in Abbildung 32 mit Hilfe der letzten beiden Zahlbeispielen zeigt. Die Konturlinien der dargestellten Mengen sind willkürlich und beliebig wählbar, insbesondere lässt die umschlossene Fläche oder die Form der umschließenden Linie keinen Schluss auf die *Mächtigkeit* (den "Umfang") der Menge zu.

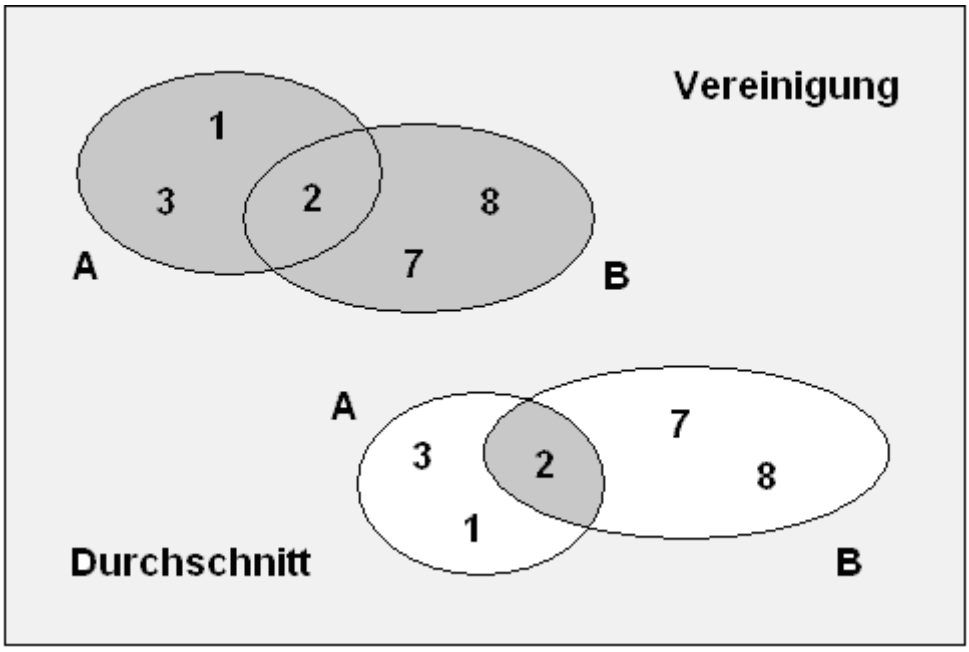


Abbildung 32: Venn-Diagramme für Vereinigung und Durchschnitt

A.3 Spezielle Symbole und Operationen

Als spezielle Symbole für Rechenoperationen werden in der Mathematik häufig das *Summenzeichen* und das *Produktzeichen* verwendet. Die Definitionen ergeben sich aus den nächsten beiden Gleichungen:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

Das *Summenzeichen* (Σ="Sigma", griech. Buchstabe für S wie Summe) wird durch die Grenzen "1" und "n" des Indexes "i" ergänzt. Damit wird festgelegt, dass der *Laufindex* i der Reihe nach die Werte 1,2,3,...,n annehmen soll.

Die Beziehung

$$\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot \ldots \cdot x_n$$

definiert in analoger Weise das *Produkt* der Zahlen x_i (Π="Pi", griechischer Buchstabe für P wie Produkt).

Die Definition einer *Potenz* erfolgt durch

$x^n = x \cdot x \cdot \ldots \cdot x$
$\leftarrow$ n-mal $\rightarrow$

wobei x n-mal mit sich selbst multipliziert wird.

Diese Sprechweise ist zwar üblich, aber nicht ganz korrekt: Im Grunde wird x nur (n-1)-mal mit sich selbst multipliziert, wie man sich am Beispiel von n=2 leicht klar machen kann.

Mit Potenzen lässt sich ohne Weiteres rechnen, wenn man die üblichen Rechenregeln beachtet. Diese Regeln sind sicher jedem noch aus der Schulmathematik bekannt und müssen hier nicht eingehend wiederholt werden. Zur Erinnerung:

1. Regel: $x^1 = x$	2. Regel: $x^0 = 1$
3. Regel: $(x \cdot y)^a = x^a \cdot y^a$	4. Regel : $(x/y)^a = x^a / y^a$
5. Regel: $x^a \cdot x^b = x^{(a+b)}$	6. Regel: $x^a / x^b = x^{(a-b)}$
6. Regel: $x^{-a} = 1/x^a$	8. Regel: $(x^a)^b = x^{a \cdot b}$

Die Gleichung $x^a=b$ bedeutet, dass die unbekannte Zahl x a-mal mit sich selbst multipliziert werden soll, um die Zahl b zu erhalten. Die Lösung dieser Aufgabe führt bekanntlich zum Begriff der *Wurzel*, und man erhält dazu als Lösung $x=\sqrt[a]{b}$. Beispiel: Die Gleichung $x^2=4$ besitzt die Lösung $x=\sqrt[2]{4}=2$. In der Regel lässt man zur Vereinfachung die Zahl 2 über dem Wurzelzeichen weg, schreibt also nur $\sqrt{4}$ und meint damit immer eine Quadratwurzel; in allen anderen Fällen ist eine Angabe verbindlich. Äquivalent zu "$\sqrt[a]{b}$" findet man gelegentlich auch die Schreibweise " $b^{1/a}$".

Ganz ähnlich ist der *Logarithmus* definiert. Betrachtet man die Gleichung $a^x=b$, so lautet hier die Frage, wie oft a mit sich selbst multipliziert werden muss, um b zu ergeben; zum Beispiel hat $10^x=1000$ die Lösung x=3. Allgemein verwendet man zur Lösung der Aufgabe den Logarithmus: Es ist $x={}_a\log(b)$. a heißt dabei Basis des Logarithmus, wobei in der Regel die Basen 2, e (Eulersche Zahl, e=2.718281….) und 10 verwendet werden. Im letzten Beispiel ist die Basis 10, und es ergibt sich als Lösung $x={}_{10}\log(1000)=3$. (Beispiel: Der ganzzahlige Anteil plus 1 von ${}_{10}\log(b)$ ergibt die Anzahl der Dezimalstellen der Zahl b!) Gelegentlich schreibt man an Stelle von ${}_2\log$ auch ld (logarithmus dualis) und an Stelle von ${}_e\log$ auch ln (logarithmus naturalis).

Beim Rechnen mit Logarithmen muss man ebenfalls einige Regeln beachten - auch hier nur zur Erinnerung:

1. Regel: $\log(1) = 0$	2. Regel: $\log(x \cdot y) = \log(x) + \log(y)$
3. Regel: $\log(x/y) = \log(x) - \log(y)$	4. Regel: $\log(x^a) = a \cdot \log(x)$
5. Regel: $\log(\sqrt[a]{x}) = \log(x) / a$	6. Regel: $\log(1/x) = -\log(x)$

Logarithmen spielen als Umkehrfunktion der Exponentialfunktion eine besondere Rolle. Als Umkehrfunktion einer Funktion f bezeichnet man eine Funktion g, für die g(f(x))=x ist; zum Beispiel ist die Quadratwurzel die Umkehrfunktion von $f(x)=x^2$: $\sqrt{x^2}=x$. Entsprechend ist $_e\log(e^x)=x$. Über Exponentialfunktionen wird erst im nächsten Abschnitt gesprochen.

In dem mathematischen Gebiet der *Kombinatorik* versteht man unter einer *Fakultät* n! (gelesen: n Fakultät) das Produkt $n!=1 \cdot 2 \cdot \ldots \cdot n$ der ersten n natürlichen Zahlen. Als Spezialfall definiert man 0!=1. Mit n! errechnet man die Anzahl Möglichkeiten ("*Permutationen*"), n verschiedene Elemente in unterschiedlicher Weise anzuordnen. Zum Beispiel gibt $3!=1 \cdot 2 \cdot 3=6$ die Anzahl verschiedener Anordnungen z.B. der Zahlen 1, 2 und 3 oder der Buchstaben a, b und c an; im letzten Beispiel also die Anzahl 6 der *Permutationen* abc, acb, bac, bca, cab und cba.

Damit erklärt sich auch die Definition von n!: Für die erste Position hat man 3 Möglichkeiten der Besetzung, für jede dieser 3 Möglichkeiten ergeben sich für die zweite Position jeweils 2 Kandidaten, bis jetzt also $3 \cdot 2=6$ Möglichkeiten. Der letzte Platz steht für jede dieser 6 unterschiedlichen Möglichkeiten fest (1 "Wahlmöglichkeit"), womit sich insgesamt also $3 \cdot 2 \cdot 1=6=3!$ unterscheidbare Anordnungen ermitteln lassen.

Sind etwa k der n Elemente gleich, so erhält man zwangsläufig weniger unterscheidbare Anordnungen, denn unabhängig davon wie man die k gleichen Elemente vertauscht, das Ergebnis ist immer das gleiche. Deshalb berechnet sich die Anzahl der *unterscheidbaren* Permutationen bei k gleichen Elementen einsichtigerweise mit n!/k!. (Beispiel: Ist k=k!=2, so kann man die Vertauschung der beiden gleichen Dinge nicht erkennen und die Anzahl unterscheidbarer Anordnungen halbiert sich.) Gibt es nun unter den n Elementen k_1 und k_2 gleiche, so muss man die Anzahl unterscheidbarer Anordnungen mit $n!/(k_1! \cdot k_2!)$ berechnen. Für $k_1=k$ und $k_2=n-k$ ergibt sich daraus der Binomialkoeffizient:

Der *Binomialkoeffizient* $\binom{n}{k}$ gibt somit die Anzahl der Möglichkeiten an, aus n verschiedenen Dingen k herauszugreifen; solche Auswahlen werden auch als *Kombinationen* bezeichnet. $\binom{n}{k}$ errechnet sich mit

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

In Abschnitt 0.5 findet sich ein Beispiel: Wieviele Möglichkeiten gibt es, unter n=4 Kindern k=2 Mädchen zu haben? Es sind $4!/(2! \cdot 2!)=6$ Kombinationen.

Betrachtet man z.B. die Menge {a,b,c}, so kann man als Paare die Mengen {a,b}, {a,c} und {b,c} bilden: Die Anzahl dieser Möglichkeiten ergibt sich mit n=3 und k=2: $\binom{n}{k}=3!/(2!\cdot 1!)=3$. Wieviele Möglichkeiten gibt es, mit k bzw. n-k jeweils gleichen Dingen (hier: k Mädchen, n-k Jungs) unterscheidbare Anordnungen herzustellen? Auch dazu verwendet man den Binomialkoeffizienten. Das sicher bekannteste Beispiel ist das Zahlenlotto: $\binom{n}{k}=13.983.816$ gibt die Anzahl Möglichkeiten an, aus n=49 Dingen (den Zahlen von 1 bis 49) k=6 bestimmte (und dies in beliebiger Reihenfolge!) herauszugreifen. Diese Überlegung geht aber schon über die Absicht dieses Abschnittes hinaus und ist eher Gegenstand der Wahrscheinlichkeitsrechnung, über die ausführlich in Kapitel 0 gesprochen wird.

Im Zahlenlotto gibt es bei der Ziehung der ersten Kugel 49 Möglichkeiten. Da diese Kugel nicht zurückgelegt wird, gibt es bei der Ziehung der zweiten Kugel nur noch 48 Möglichkeiten, bis jetzt also 49·48 mögliche Zahlenpaare. Die dritte Kugel wird aus den restlichen 47 Kugeln ausgewählt, die vierte aus 46, die fünfte aus 45 und die sechste aus 44: Insgesamt ergeben sich also 49·48·47·46·45·44 verschiedene Zahlenfolgen. Diese Anzahl kann man auch ausdrücken durch n!/(n-k)!=49!/(49-6)!=49!/43!. Da im Lottospiel die Reihenfolge der gezogenen k=6 Zahlen keine Rolle spielt, also jede beliebige Permutation der gezogenen k=6 Zahlen gleichwertig ist, dividiert man n!/(n-k)! durch die Größe k!=6! und erhält die Anzahl der unterschiedlichen 6-elementrigen Teilmengen von {1,2,...,49} bzw. die Anzahl der möglichen, unterschiedlichen Lottoergebnisse mit n!/(k!·(n-k)!), per definitionem also den Binomialkoeffizienten.

A.4 Funktionen

Eine mathematische *Funktion* ist definiert als eine *eindeutige* Zuordnung einer Menge ***X*** zu einer Menge ***Y***. Bekannte Beispiele sind Geradengleichungen (y=a+b·x), *Parabeln* (y=a+b·x+c·x², speziell die *Normalparabel* y=x²) oder die *Exponentialfunktion* (in ihrer speziellen Form $y=a^x$), die alle eine solche "Zuordnungvorschrift" definieren. Eine Eindeutigkeit in umgekehrter Richtung muss nicht zwangsläufig gegeben sein, wie bereits das Beispiel der Normalparabel zeigt (Inverse Funktion, es ist $x=\pm\sqrt{y}$). Beispiele hierzu sind jedem geläufig, wenn man die Mengen ***X*** und ***Y*** mit der Menge der reellen Zahlen ***R*** identifiziert.

Betrachtet man eine Gerade, die durch den *Nullpunkt* (den "*Ursprung*") geht, so folgt diese der Beziehung y=b·x. Löst man diese Gleichung nach b auf, so erhält man b=y/x. Dieser Quotient wiederum gibt die Steigung der Geraden an. Eine weitere Interpretation erhält man, wenn man sich an die trigonometrischen Beziehungen erinnert, die in rechtwinkligen Dreiecken gelten: Mit dem Anstiegswinkel α der Geraden, y als Gegenkathete und x als Ankathete kann man b auch definieren durch $b=\tan(\alpha)=y/x$. Möchte man sich nicht nur auf den Ursprung beziehen und auch Geraden mit $a\neq 0$ betrachten (a ist die Stelle auf der y-Achse, der Ordinate, an der die

Gerade bei x=0 die y-Achse schneidet, "*Achsenabschnitt*"), so ergibt sich ganz analog eine Definition der Steigung b als der *Differenzenquotient* $b=(y_2-y_1)/(x_2-x_1)=\Delta y/\Delta x$.

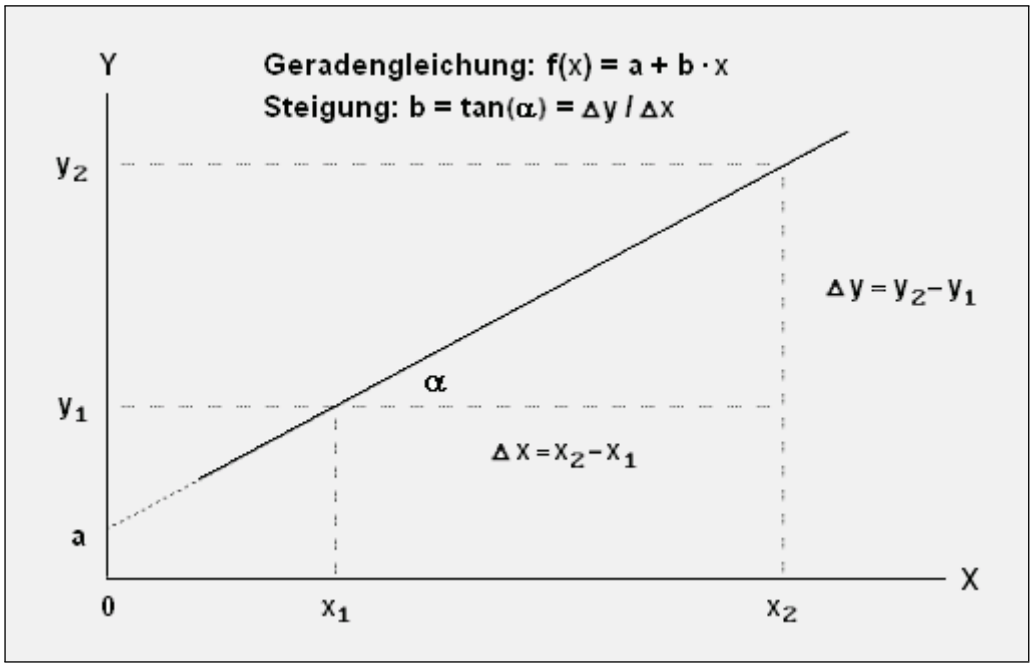


Abbildung 33: Definitionen zur Geraden

Als einfache quadratische Funktion ist jedem die *Normalparabel* geläufig, die definiert ist mit $y=x^2$. In allgemeinerer Form beschreibt man eine Parabel durch $y=a+b\cdot x+c\cdot x^2$; die "*Koeffizienten*" a, b und c sind - der Reihe nach - dem *Achsenabschnitt*, der *linearen Komponente* x und der *quadratischen Komponente* x^2 zugeordnet. Bei einer quadratischen Funktion ohne Weiteres von einer Steigung zu reden, ist sicher nicht so einfach wie im Fall einer Geraden. Darauf wird später wieder zurückgekommen. Eine graphische Darstellung einer Parabel findet man in Abbildung 36.

Die zuletzt beschriebene Funktionsgleichung $y=x^2$ der Normalparabel stellt eine spezielle Form eines Polynoms dar. Polynome sind definiert durch $y=b_o+b_1\cdot x+b_2\cdot x^2+....+b_n\cdot x^n=\Sigma_i(b_i\cdot x^i)$, wobei von i=0 bis i=n summiert wird ($x^o=1$!). Man spricht auch von *Polynomen n-ten Grades* (höchste Potenz!), insbesondere ist eine Parabel ein Polynom 2. Grades.

Exponentialfunktionen trifft man in der Medizin in vielen Bereichen an. Diese sind definiert durch die Beziehung $y=a\cdot e^{bx}$, oft auch geschrieben $y=a\cdot\exp(b\cdot x)$ und werden zum Beispiel zur Beschreibung des radioaktiven Zerfalls, bei Untersuchung der Absorption von Licht, bei allen Wachstumsprozessen etc. benötigt. So wird etwa das Gesetz zur Beschreibung des radioaktiven Zerfalls durch die Gleichung $C_t=C_0\cdot\exp(-\lambda\cdot t)$ beschrieben:

Dabei ist C_0 die Aktivität zum Zeitpunkt 0 (i. a. Versuchsbeginn), t ist ein beliebiger Zeitpunkt nach Versuchsbeginn, $C(t)=C_t$ ist die Aktivität zu einem Zeitpunkt t, λ ist die spezifische Zerfallskonstante.

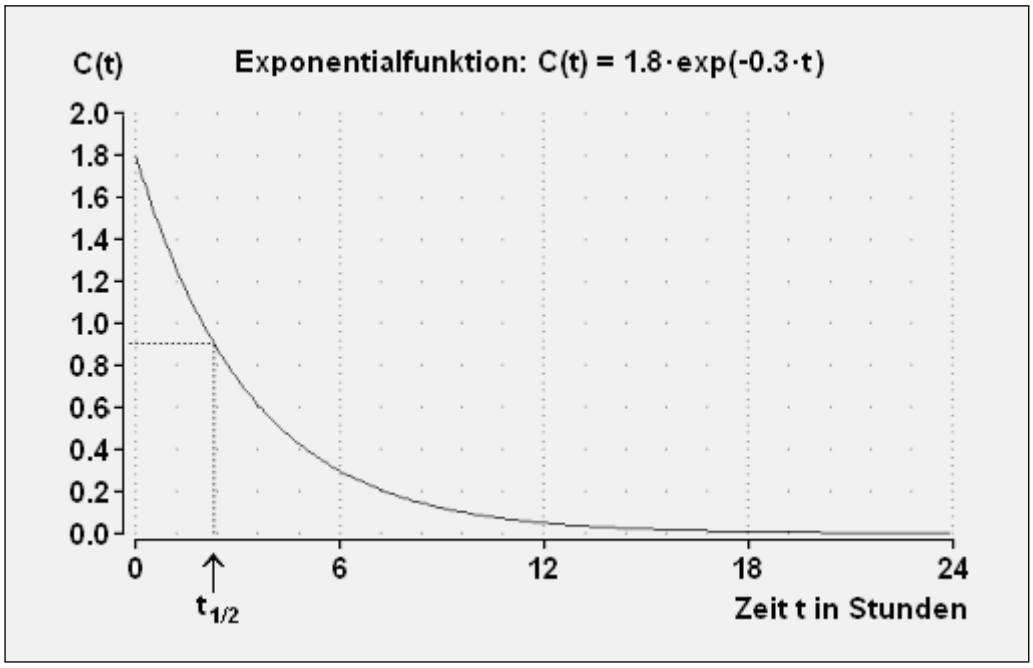


Abbildung 34: Exponentialfunktion

Die in Abbildung 34 dargestellte Exponentialfunktion $C(t)=C_t=1.8\cdot\exp(-0.3\cdot t)$ beschreibt die Aktivität radioaktiven Materials über 24h. Aus der Funktion kann man leicht die Halbwertszeit $t_{1/2}$ berechnen, die einfacher zu interpretieren ist als die Zerfallskonstante λ=0.3: Die Halbwertszeit $t_{1/2}$ ist die Zeitspanne, nach der nur noch die Hälfte des gegebenen Materials vorhanden bzw. aktiv ist. Aus dem Ansatz $C_0/2 = C_0\cdot\exp(-\lambda\cdot t_{1/2})$ ergibt sich nach Logarithmierung der Gleichung (vgl. Rechenregeln in Anhang A.3!) $\log(C_0)-\log(2)=\log(C_0)-(\lambda\cdot t_{1/2})$, hieraus $-\log(2)=-\lambda\cdot t_{1/2}$ und deshalb $t_{1/2}=\log(2)/\lambda$. Im Beispiel ist $t_{1/2}=\log(2)/0.3=2.3105$.

Spezielle Exponentialfunktionen, sogenannte *Bateman-Funktionen*, finden sich häufig bei der pharmakokinetischen Beschreibung von *Konzentration-Zeit-Verläufen* der Serumkonzentration von Pharmaka. Abbildung 35 verschafft einen Eindruck über den typischen Verlauf solcher Kurven.

Die Serum-Konzentration-Zeit-Verläufe in Abbildung 35 ergeben sich als Summen von zwei Exponential-Funktionen im Zusammenspiel der Absorption und Elimination eines Pharmakons (Aufnahme und Abbau eines Medikamentes im Serum, hier von Methadon). Im vorliegenden Modell der oralen Applikation ist die Serumkonzentration berechenbar aus $C(t)=C_0\cdot k_a/(k_a-k_e)\{\exp(-k_e\cdot t)-\exp(-k_a\cdot t)\}$ mit den Geschwindigkeitskonstanten k_a=Absorptionskonstante und k_e=Eliminationskonstante ("Stoffumsatz pro Zeit", Halbwertszeiten berechnen wie oben!). Die Kurve ohne "Zacken" ergibt sich als Verlauf der Serumkonzentration bei täglich einmaliger Methadon-Gabe zum Zeitpunkt t_0=0h, die zweite Kurve als Konzentrationsverlauf bei täglich zweimaliger, jedoch jeweils halbierter Dosis der Wirksubstanz zu den Zeitpunkten t_0=0h und t=12h. Es zeigt sich die höhere Effektivität der zweimaligen Gabe,

die einen höheren "Talspiegel" und damit geringere Entzugserscheinungen mit sich bringt, dazu Hellenbrecht, Ackermann, Saller 1994, Tägl. Praxis 35, pp. 477ff.

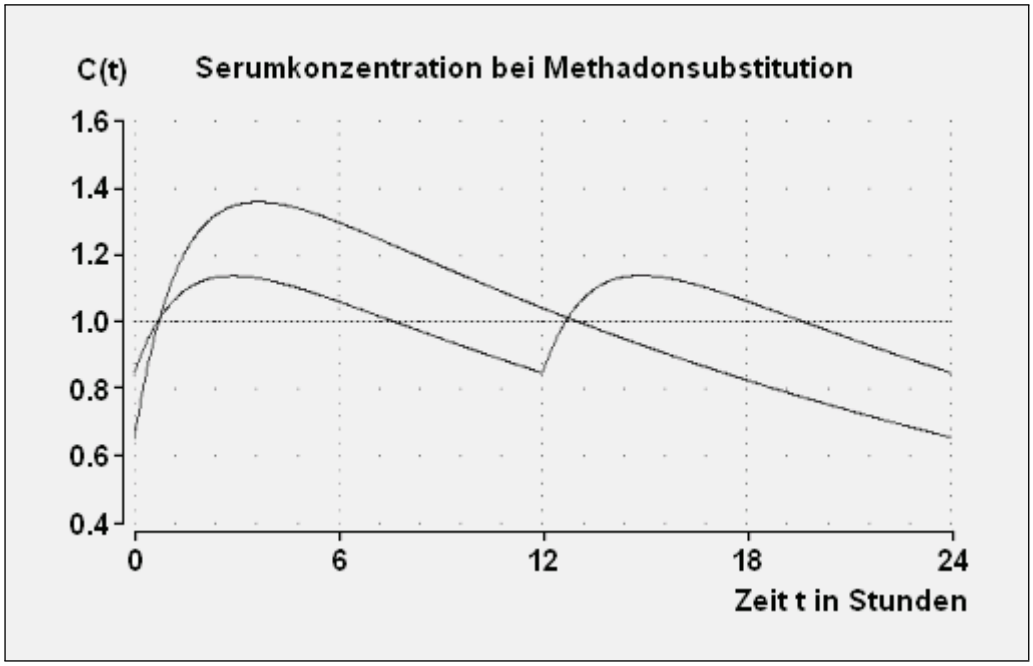


Abbildung 35: Konzentration-Zeit-Verläufe t versus C(t) nach Medikation

A.5 Differentialrechnung

Im letzten Abschnitt wurde bereits festgestellt, dass eine Angabe einer "Steigung" im Falle der Parabel (natürlich auch bei Polynomen oder anderen, komplizierteren Funktionen) sicher nicht einfach ist, denn eine gekrümmte Kurve weist offenbar an jeder Stelle eine andere "Steigung" auf. Um sich dem Problem anzunähern, betrachtet man zwei beliebige Werte $X=x_1$ und $X=x_2$ zusammen mit deren Funktionswerten $Y=y_1=f(x_1)$ bzw. $Y=y_2=f(x_2)$, wie aus Abbildung 36 ersichtlich ist.

Durch die beiden Punkte (x_1,y_1) und (x_2,y_2) kann man nun eine Gerade legen (die auch als *Sekante* bezeichnet wird) und deren Steigung $b=\Delta y/\Delta x=(y_2-y_1)/(x_2-x_1)$ bestimmen, die natürlich nur "ungefähr" mit der Steigung der Kurve übereinstimmt. lässt man nun Δx immer kleiner werden (x_2-x_1 bzw. $\Delta x \rightarrow 0$), so verbessert sich die Situation immer mehr und man erhält "im Grenzfall" die Steigung der Kurve in dem Punkt x_1, die mit der Steigung der *Tangente* in x_1 identisch ist. Der Mathematiker Leibnitz (1646-1710) konnte zeigen, dass diese Grenzwertbildung (bei der Δx nicht identisch Null wird!) für viele Funktionen sinnvolle Ergebnisse liefert: Diese Funktionen werden als "*differenzierbar*" bezeichnet. Den

Prozess, Δx beliebig klein zu machen, bezeichnet man auch als "*Grenzwertbildung*" (Limes, abgekürzt lim), in einer Formel ausgedrückt

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \frac{dy}{dx}$$

Den Quotienten "dy/dx" bezeichnet man als *Differentialquotienten* oder auch als *Ableitung* und kürzt dies häufig ab mit y' oder f'(x).

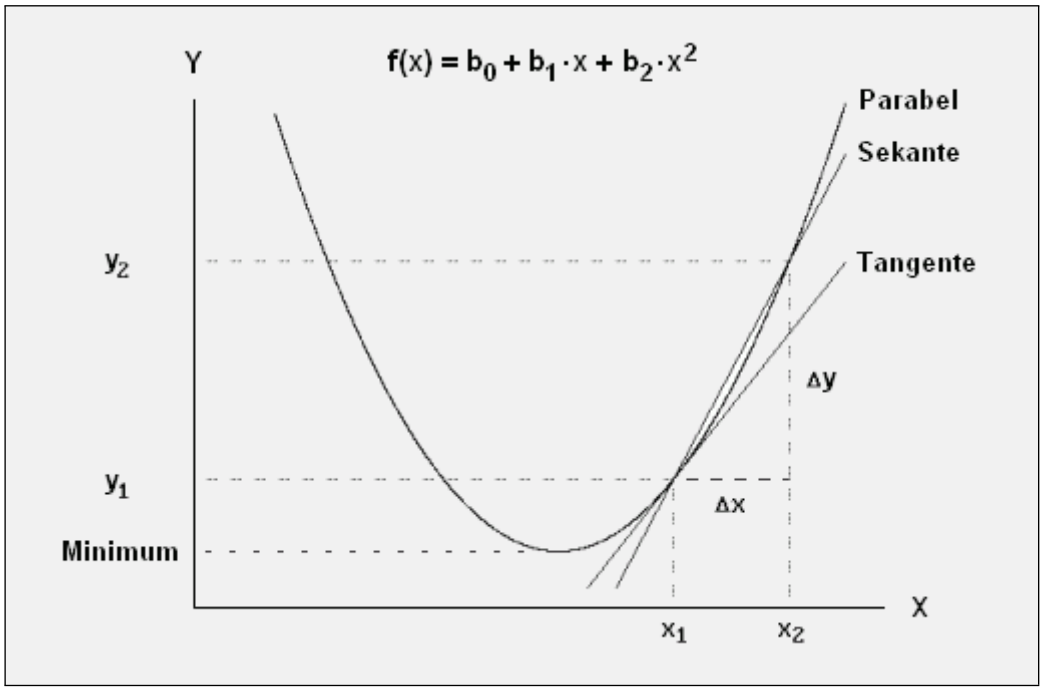


Abbildung 36: Definitionen zur Differentialrechnung

Die Ableitung eines Polynoms ist recht einfach herzustellen, wenn man die Terme $b_i \cdot x_i$ ($i>0$) untersucht: Deren Ableitung ist $i \cdot b_i \cdot x^{i-1}$. Die Ableitung des Polynoms ergibt sich nun als Summe der Ableitungen der n einzelnen Summenterme. Die Ableitung des konstanten Gliedes b_o (dies hat keinen Term mit x!) ergibt Null.

Allgemein bekannt ist die Ableitung für die Normalparabel $f(x)=y=x^2$, die mit der Beziehung $f'(x)=y'=2 \cdot x$ angegeben werden kann. An derjenigen Stelle $x=x_0$, an der die Ableitung (i.e. Steigung!) der Funktion Null ist, liegt der Extremwert der Parabel (aus $y'=2 \cdot x=0$ erhält man $x=x_0=0$). Dies kann ein *Maximum* oder ein *Minimum* sein. Aufschluss erhält man über die 2. Ableitung der Funktion als Ableitung der 1. Ableitung; diese wird auch als Krümmung bezeichnet. Ist die 2. Ableitung an der Stelle x_0 positiv (im Beispiel ist $y''=2>0$ für alle x), so schließt man auf ein Minimum an der Stelle $x=x_0$, ist die 2. Ableitung dagegen negativ, so muss an der Stelle $x=x_0$ ein Maximum vorliegen.

Die erste Ableitung einer quadratischen Funktion spielt in der Regressionsrechnung eine besondere Rolle (Abschnitt 5.7, "Methode der kleinsten Quadrate"). Die fragliche Funktion ist dort eine Summe von quadrierten Abständen einer zweidimensionalen "Punkteverteilung" von einer Geraden, dies mit dem Ziel, eine "optimale" Gerade - die Regressionsgerade - zu erhalten. Gesucht ist dabei das Minimum der erwähnten Summe im Sinne einer "besten" Beschreibung der Abhängigkeit zweier biologischer Größen (z.B. die Körperlänge in Abhängigkeit vom Alter). Die Regressionsgerade besitzt - als Gerade - ihrerseits eine Steigung, die im Sinne von Anhang A.4 interpretiert werden kann: Um welchen Betrag verändert sich die Körperlänge pro Lebensjahr?

Die Ableitung einer Exponentialfunktion ist ebenfalls leicht zu bilden. Für die Funktion $f(x)=y=a\cdot\exp(b\cdot x)=a\cdot e^{b\cdot x}$ ergibt sich die erste Ableitung $f'(x)$ mit $f'(x)=y'=a\cdot b\cdot\exp(b\cdot x)$. Im speziellen Fall, das heißt für $a=b=1$, ist $f(x)=\exp(x)=e^x$ und die erste Ableitung ist $f'(x)=f(x)=e^x$. Die Ableitungen von trigonometrischen Funktionen (sin, cos, tan etc.) sind im Allgemeinen nur etwas umständlicher zu erhalten, spielen aber in diesem Skriptum glücklicherweise keine weitere Rolle.

Im Beispiel der Bateman-Funktionen (Abbildung 35) interessiert man sich in der Pharmakologie für den Zeitpunkt t_{max} der maximalen Konzentration und für die maximale Konzentration C_{max}. Bildet man die 1. Ableitung, setzt diese gleich Null und löst die Beziehung nach t auf, so erhält man ein Extremum der Funktion, das hier nur ein Maximum sein kann: Das Maximum liegt bei $t=t_{max}$. Setzt man diesen Wert t_{max} in die Funktionsgleichung ein, so ergibt sich daraus der Wert $C_{max}=C(t_{max})$.

A.6 Integralrechnung

Ein *Integral* kann man als diejenige Fläche auffassen, die in einem Bereich x_1 bis x_2 ($x_1<x_2$) von einer Kurve und der Abszisse ("X-Achse") eingeschlossen wird. Bei einer Geraden als Funktion ist die Berechnung ganz sicher kein Problem, denn jeder weiß, wie man die Fläche eines Dreiecks auszurechnen hat. Schwieriger wird es also wieder, wenn kompliziertere Funktionen als nur Geraden untersucht werden.

Das "Integral" der *Winkelhalbierenden* $f(x)=y=x$ kann man ohne jede Kenntnis der Integralrechnung bestimmen, da sich hier das Problem auf die Berechnung der Fläche eines Dreiecks reduziert. Zum Beispiel berechnet sich für $f(x)=x$ die "Fläche unter der Kurve" zwischen 0 und x, etwa $x=5$, bekanntlich mit $0.5\cdot x^2=0.5\cdot 5^2=0.5\cdot 25=12.5$. Im vorletzten Absatz dieses Abschnittes erhält man das gleiche Resultat etwas komplizierter und formal über die sogenannte *Stammfunktion* $F(x)$ der Funktion $f(x)=x$.

Für eine beliebige Funktion $y=f(x)$ - zum Beispiel für ein Polynom, eine Exponentialfunktion o.a. - soll nun diejenige Fläche berechnet werden, die zwischen den Abszissenwerten $x=x_1$ und $x=x_2$ von der Kurve und der Abszisse eingeschlossen wird. Dazu zerlegt man die Strecke x_1 bis x_2 in n gleichlange Teilstückchen, jedes also mit der Länge $\Delta x=(x_2-x_1)/n$. Da sich innerhalb jedes dieser kleinen Teilstückchen die Kurve nicht wesentlich

verändert, kann man die interessierende Fläche A angenähert als Summe von n schmalen Rechtecken berechnen:

$$A = \sum_{i=1}^{n} f(x_1 + i \cdot \Delta x) \cdot \Delta x$$

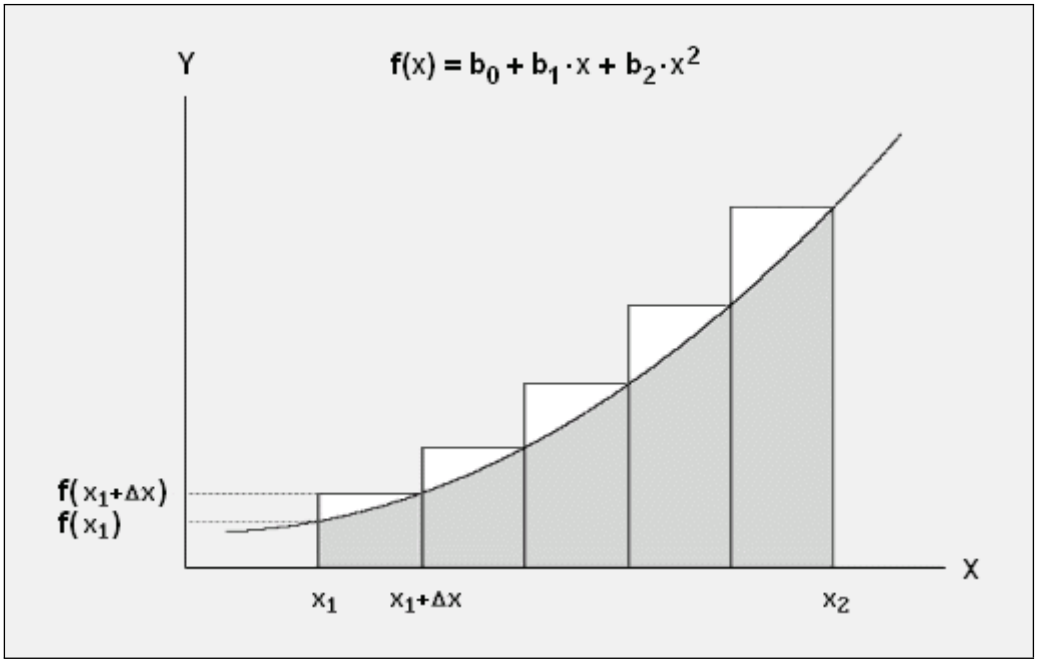


Abbildung 37: Definitionen zur Integralrechnung

An Stelle der „oberen“ Summe $A=\Sigma f(x_1+i\cdot\Delta x)\cdot\Delta x$ wie oben könnte man auch analog die „untere“ Summe $A'=\Sigma f(x_1+(i-1)\cdot\Delta x)\cdot\Delta x$ bilden, die jedoch für $\Delta x\rightarrow 0$ offensichtlich zum gleichen Resultat führt.

Natürlich ist die Berechnung von A umso genauer, je kleiner Δx ist. Somit bietet sich, wie im letzten Abschnitt, ein Grenzübergang $\Delta x\rightarrow 0$ bzw. $n\rightarrow\infty$ an, eine Überlegung, die ebenfalls auf den bereits zitierten Mathematiker Leibnitz zurückgeht (dies ist der sogenannte "*Hauptsatz der Differential- und Integralrechnung*"):

$$\lim_{\Delta x \to 0} A = \int_{x_1}^{x_2} f(x)dx = F(x) \Big|_{x_1}^{x_2} = F(x_2) - F(x_1)$$

F(x) heißt *Stammfunktion* oder auch *unbestimmtes Integral* der Funktion f(x); Integrale mit festen Integrationsgrenzen (wenn also letztendlich konkrete Zahlenwerte vorliegen) bezeichnet man als *bestimmte Integrale*. Die Bestimmung von Integralen ist unter Umständen recht einfach (zum

Beispiel bei Polynomen), unter Umständen aber auch recht kompliziert oder – in expliziter Form - sogar unmöglich (vgl. Abschnitt 4.2 zur Gauß-Verteilung), so dass man sich gegebenenfalls anspruchsvollerer Methoden aus der *Numerischen Mathematik* bedienen muss.

Setzt man oben für f(x) die Geradengleichung f(x)=y=x ein, so erhält man die Stammfunktion $F(x)=x^2/2+const.$. Die Konstante "const." ist unbekannt, spielt aber bei der Berechnung des Integrals (vgl. oben, bestimmtes Integral!) wegen der Differenz $F(x_2)-F(x_1)$ offenbar keine Rolle. Diese Zusammenhänge werden, wie bereits erwähnt, im *Hauptsatz der Differential- und Integralrechnung* behandelt.

Allgemein ist das Integral eines Polynomterms $b_i \cdot x_i$ mit $(1/(i+1)) \cdot b_i \cdot x^{i+1}+$ +const. gegeben, so dass das Integral eines vollständigen Polynoms via Summenbildung ohne Weiteres berechnet werden kann. Diese Beziehung ist leicht einzusehen, denn bekanntlich besteht zwischen der Stammfunktion F(x) und der zu integrierenden Funktion f(x) ein sehr enger Zusammenhang: Bildet man die erste Ableitung der Stammfunktion, so erhält man den Integrand f(x): Es ist F'(x)=f(x). Dieses Wissen genügt, um bei relativ einfachen Funktionen - wie z.B. bei Polynomen - im konkreten Fall das Integral zu ermitteln. In schwierigeren Fällen können umfangreiche Tabellen oder auch einschlägige Computerprogramme (im Shareware-Bereich Programme wie "MatheAss" oder "Mercury") verwendet werden.

Im Beispiel der Bateman-Funktion (Abschnitt A3, Abbildung 35) ist das Integral pharmakokinetisch von besonderem Interesse. Bei der "Fläche unter der Kurve" spricht man in der Pharmakologie von der AUC ("Area Under Curve") und versteht diese als Maß für die sog. *Bioverfügbarkeit* der verabreichten Wirksubstanz: Eine Substanz kann als umso effektiver aufgefasst werden, je größer ihr AUC-Wert ist. In der Praxis liegen allerdings nur z.B. n Messwerte-Paare $(t_i,C(t_i))$ vor, so dass man in der Regel an Stelle des Integrals den entstehenden Polygonzug untersucht und die Fläche unter diesem Polygonzug rein geometrisch mit Rechtecken und Dreiecken bestimmt. Neben der AUC verwendet man in der Pharmakologie häufig zusätzlich die charakteristischen Größen t_{max} (Zeitpunkt der maximalen Serumkonzentration) und C_{max} (maximale Serumkonzentration).

Der sogenannte *Talspiegel* (Minimum der Konzentration im Zeitverlauf, C_{min}) ist im Beispiel der Abbildung 35 aus medizinischer Sicht ebenfalls von besonderem Interesse: Die beiden Bioverfügbarkeiten bzw. die beiden AUC's als Integrale für die volle und die zweimal-halbierte Dosierung sind identisch, der Talspiegel im zweiten Fall ist dagegen deutlich größer als im ersten. Auf eine medizinische Interpretation wurde bereits weiter oben in Anhang A.4 hingewiesen.

Das Programmpaket **BiAS. für Windows** berechnet alle eben genannten Größen, daneben auch aus der Bateman-Funktion bzw. aus dem Integral abgeleitete Größen wie die AUC von t_1 bis t_n, die AUC von t_1 bis ∞, die Terminale Eliminationskonstante λ, die Terminale Halbwertszeit $t_{1/2}$ vermittels loglinearer Regression mit den terminalen Zeitpunkten, die Mittlere Konzentration C_{av}, die prozentuale Peak-Through-Fluktuation PTF, den prozentualen Swing PTS, die Half-Value-Duration HVD, die Mean-Residence-Time MRT und die AUCM(t_1-∞) als Fläche unter der 1. Momentkurve.

Literatur

Abel U (1993) Die Bewertung diagnostischer Tests. Hippokrates-Verlag Stuttgart.

Ackermann H (1989-2019) BiAS.: Biometrische Analyse von Stichproben (Programmpaket). Version 11, epsilon-Verlag https://www.bias-online.de

Ackermann H (1994) Medizinische Normbereiche. Med Welt 45, 11, pp. 448-56.

Ackermann H, Herrmann E (2014) Epidemiologie. In: Alles fürs Examen: Kompendium für die 2. Ärztliche Prüfung. 2. Aufl. Thieme-Verlag, Seiten 834-40. Siehe auch: Herrmann E, Ackermann H (2014) Medizinische Biometrie, in AllEx, Thieme-Verlag.

Altman DG (1998) Statistical Reviewing for Medical Journals, Stat. in Med. 17, 1998, pp. 2661-2674.

Altman DG (Editor, 1994-2007) Statistical Notes. Brit. Med. J. http://www.bmj.com und http://www.jerrydallal.com/LHSP/bmj.htm

Armitage P, Berry G (1988,2002) Statistical Methods in Medical Research. Blackwell Scientific Publications, 2nd edition.

Bauer P, Köhne K (1994) Evaluation of experiments with adaptive interim analysis. Biometrics 50, pp. 1029-1041.

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurements. Lancet 8.2.1986.

Bundesärztekammer (2008) Richtlinien zur Qualitätssicherung laboratoriumsmedizinischer Untersuchungen. Dtsch. Ärzteblatt Jg. 105, Heft 7.

Bundesinstitut für Arzneimittel und Medizinprodukte BfArM (1998) Empfehlungen zur Planung und Durchführung von Anwendungsbeobachtungen. Bundesanzeiger Jg. 50, 229, Seite 16884

Ciba-Geigy (Hrsg., 1985) Wissenschaftliche Tabellen Geigy, Teilband Statistik 8. Auflage, Ciba-Geigy AG, Basel

D'Agostino RB Jr (1998) Tutorial in Biostatistics: Propensity Score methods for bias reduction in the comparison of a treatment to a non randomized control group. Statistics in Medicine 17, pp. 2265-2281.

Dallal GE (2000,2012) The little handbook of statistical practice. Tufts University Boston, MA 02111 und http://www.jerrydallal.com/LHSP/LHSP.HTM

Deichsel G, Trampisch HJ (1985) Clusteranalyse und Diskriminanzanalyse. Gustav Fischer Stuttgart New York.

Fleiss JL, Levin B, Paik MC (2003) Statistical Methods for Rates and Proportions. Wiley Series in Probability and Statistics.

Fletcher RH, Fletcher SW (2007) Klinische Epidemiologie. 2. Aufl. Hans Huber, Bern.

Guggenmoos-Holzmann I, Wernecke KD (1996) Medizinische Statistik. Blackwell Wissenschaftsverlag Berlin Wien.

Held L, Rufibach K, Seifert B (2013) Medizinische Statistik. Pearson Higher Education München.

Herrmann E, Ackermann H (2014) Medizinische Biometrie. In: Alles fürs Examen: Kompendium für die 2. Ärztliche Prüfung. 2. Auflage Thieme-Verlag, Seiten 840-852.

Hilgers RD, Bauer P, Scheiber V (2003) Einführung in die Medizinische Statistik. Springer-Verlag Berlin Heidelberg New York.

Hollander M, Wolfe DA (1999) Nonparametric Statistical Methods. Wiley Series in Probability and Statistics, 2nd edition.

Horn M, Vollandt R (1995) Multiple Tests und Auswahlverfahren. Gustav Fischer - Verlag Stuttgart-Jena.

ICH/GCP: The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (1990-2005), Genf/CH.

Kaatsch P, Spix C, Jung I, Blettner M (2008) Childhood Leukemia in the Vicinity of Nuclear Power Plants in Germany. Dtsch. Arzteblatt Int. 105 (42): pp. 725–732

Kabisch M, Ruckes C, Seibert-Grafe M, Blettner M (2011) Randomized Controlled Trials. Dtsch. Ärzteblatt Int. 108 (39), pp. 663-668

Klein JP, Moeschberger ML (1997) Survival-Analysis. Techniques for censored and truncated data. Springer New York.

Kreienbrock L, Schach S (2005) Epidemiologische Methoden. 3. Auflage Elsevier.

Lautsch E, Lienert LA (1993) Binärdatenanalyse. Beltz-Verlag.

Lorenz RJ (1996) Grundbegriffe der Biometrie. 4. Aufl. Gustav Fischer-Verl. Stuttgart

Mace AE (1964) Sample-Size Determination. Chapman & Hall, Ltd., London

Mantel N (1970) Why stepdown procedures in variable selection. Technometrics 12, 3, pp. 621-625.

Meyer FP (2008) Nicht-Unterlegenheitsstudien: Fragwürdige Ethik. Dtsch. Ärzteblatt 2008, 105(43) A2268-9

O'Brien PC, Fleming TR (1979) A multiple testing procedure for clinical trials. Biometrics 35, pp. 549-56.

Pocock SJ (1977) Group sequential methods in the design and analysis of clinical trials. Biometrika 64, pp. 191-199.

Sachs L (1967-2018) Angewandte Statistik. 1.-16. Auflage Springer-Verlag Heidelberg New York (ab der 12. Auflage 2009 mit Coautor Jürgen Hedderich als „Angewandte Statistik - Methodensammlung mit R")

Scheid H (2012) Abiturwissen Mathematik. 6. Auflage Ernst Klett-Verlag München.

Schneider B (2001) Beobachtungsstudien als Mittel der Erkenntnisgewinnung über die Wirksamkeit von Arzneimittel. Preprint Med. Hochschule Hannover http://www.mh-hannover.de/fileadmin/institute/biometrie/Scripte/speziel/beob_studien.pdf

Trampisch HJ, Windeler J (2000) Medizinische Statistik. 2. Auflage Springer-Verlag Berlin Heidelberg New York.

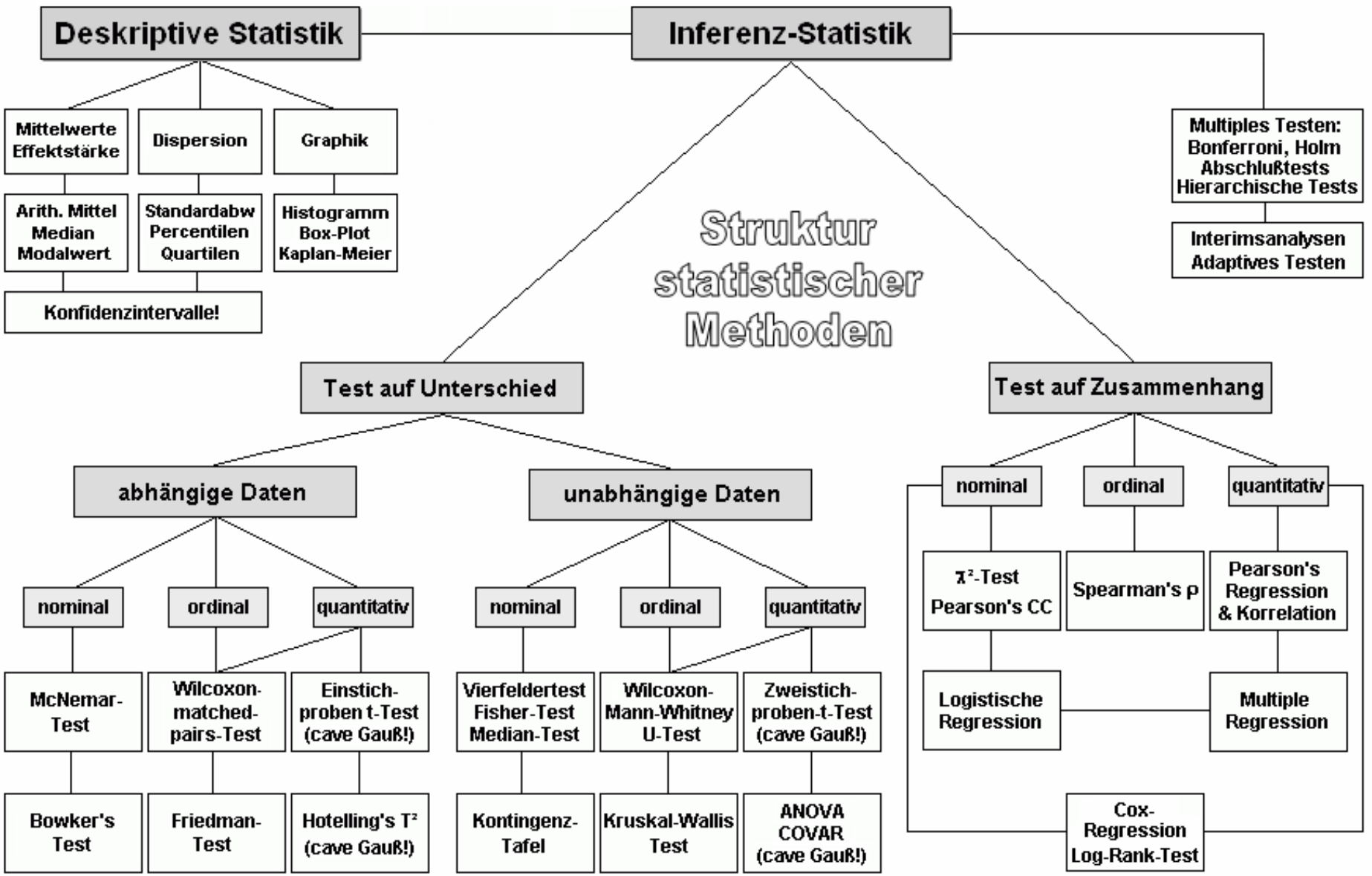
Weiß C (2013) Basiswissen Medizinische Statistik. 6. Auflage Springer-Verlag Berlin Heidelberg New York.

Wellek S (1994) Statistische Methoden zum Nachweis von Äquivalenz. Gustav Fischer - Verlag Stuttgart - Jena - New York.

Werner J (1992) Biomathematik und Medizinische Statistik. 2. Auflage Urban und Schwarzenberg München-Wien-Baltimore.

Zwiener I, Blettner M, Hommel G (2011) Survival Analysis. Dtsch. Arztebl. Int. 108, 10, pp. 163-169

Sachverzeichnis


Deskriptive Statistik
Inferenz-Statistik
Mittelwerte Effektstärke
Dispersion
Graphik
Arith. Mittel Median Modalwert
Standardabw Percentilen Quartilen
Histogramm Box-Plot Kaplan-Meier
Konfidenzintervalle!
Multiples Testen: Bonferroni, Holm Abschlußtests Hierarchische Tests
Interimsanalysen Adaptives Testen
Struktur statistischer Methoden
Test auf Unterschied
Test auf Zusammenhang
abhängige Daten
unabhängige Daten
nominal
ordinal
quantitativ
nominal
ordinal
quantitativ
nominal
ordinal
quantitativ
McNemar-Test
Wilcoxon-matched-pairs-Test
Einstich-proben t-Test (cave Gauß!)
Vierfeldertest Fisher-Test Median-Test
Wilcoxon-Mann-Whitney U-Test
Zweistich-proben-t-Test (cave Gauß!)
χ²-Test Pearson's CC
Spearman's ρ
Pearson's Regression & Korrelation
Logistische Regression
Multiple Regression
Bowker's Test
Friedman-Test
Hotelling's T² (cave Gauß!)
Kontingenz-Tafel
Kruskal-Wallis Test
ANOVA COVAR (cave Gauß!)
Cox-Regression Log-Rank-Test

Links

Informationsschrift der Biometrischen Gesellschaft (Download Pdf 1.2Mb)
http://www.dkfz-heidelberg.de/biostatistics/IBS/Biometriefolder_Internet.pdf

Deutsche Region der "International Biometric Society"
http://www.biometrische-gesellschaft.de/

Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS)
http://www.gmds.de/

Guidelines: Biostatistical Methodogy in Clinical Trials (Eudralex)
http://pharmacos.eudra.org/F2/eudralex/vol-3/pdfs-en/3cc7aen.pdf

ICH-Guidelines "Statistical Principles for Clinical Trials" (Topic E9, Emea)
http://www.ich.org/MediaServer.jser?@_ID=485&@_MODE=GLB

Guidelines for Good Clinical Practice ("GCP-Richtlinien", ICH-Topic E6)
http://www.eortc.be/Services/Doc/ICH_GCP.pdf

Empfehlungen der DFG zur Sicherung Guter Wissenschaftlicher Praxis ("GSP")
http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

Allgemeines vom Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM)
http://www.bfarm.de/

Fragen zu den gesetzlichen Bestimmungen im Arzneimittelgesetz (BfArM)
http://www.bfarm.de/de/Arzneimittel/klin_pr/klin_pr_faq/index.php

Elektronische Volltext-Zeitschriften der Uni Regensburg
http://rzblx1.uni-regensburg.de/ezeit/fl.phtml?notation=WW-YZ&bibid=UBHOH&frames=&colors=7&SC=R

Literatursuche: Entrez PubMed: National Library of Medicine
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?SUBMIT=y

Medline.de: Suchmaschine für mehr als 2000 medizinische Fachzeitschriften
http://www.medline.de/

DIMDI: Deutsches Institut für Medizinische Dokumentation und Information
http://www.dimdi.de/dynamic/de/index.html

Aktuelle Zeitschriften in Kurzfassung bei MedAustria
http://www.medaustria.at/

Lernmaterialien "Neue Statistik - eLearning"
http://www.neuestatistik.de/

JUMBO: Biometrisches Online-Lehr- und Lernsystem der Univ. Münster
http://imib.uni-muenster.de/fileadmin/template/conf/imib/lehre/skripte/biomathe/jumbo.html

Als Medizinstudent erfolgreich promovieren (via medici online)
http://www.thieme.de/viamedici/medizinstudium/promotion/uebersicht.html

Statistikprogramm: BiAS. für Windows
https://www.bias-online.de

Excel und Statistik
https://www.practicalstats.com/xlstats/excelstats.html

BiAS. für Windows *TM*

Datenmanagement
Editor, Daten-Transformationen, Formelinterpreter, Import und Export von Excel-, dBase-, SPSS-, ASCII- und Windows-Dateien (XLS, DBF, SAV, ASC, CSV, TXT)

Datenbankmodul
Selektion, Projektion, Addition, Relationieren, Sortieren, Klassifizieren (auch mit χ^2-Test), Zählen.

Fallzahl- und Powerberechnung
Konfidenzintervalle, Ein- und Zweistichproben-Tests, Varianzanalyse, Kruskal-Wallis, Schätzung von Wahrscheinlichkeiten, Vergleich binomialer Größen, Regression und Korrelation, Cross-Over, Log-Rank-Test, Äquivalenztests, Many-One-Vergleiche, Epidemiologie (OR und RR etc.), Bland-Altman-Methodenvergleich, Bauer-Köhne u.v.a..

Randomisierung
Vollständige Randomisierung, Randomisierung in Blöcken, Cross-Over-Versuchspläne, Lateinische Quadrate, Zufallsstichproben aus Gleich- und Normalverteilung.

Deskriptive Statistik
Durchschnitt, Standardabweichung, Extremwerte, Median, Quartilen, Variationskoeffizient, viele Graphiken.

Graphik
Histogramme (stetig, diskret, kumulativ, vergleichend), Kernschätzer, Box- und Bar-Plot, Kreisdiagramme, Pareto, XY-Plot, Zeitreihen-Plot, Gauß-Anpassung, Regressionsfunktion mit Konfidenzintervall, Hahn-Prognose-Intervall, Konfidenzintervalle, Kaplan-Meier-Plot, Cluster- und Diskriminanzanalyse, Bland-Altman-Plot, ROC-Kurven, GIF, JPG, PCX und BMP-Graphik-Export, auch per Clipboard.

Konfidenzintervalle
Konfidenzintervalle für μ und μ_1-μ_2, Regression und Korrelation, Median (u.a. Moses- und Tukey-Intervalle), Survival-Analyse (Intervalle nach Hall und Wellner), Simultane Konfidenzintervalle (Hotelling), Binomial- und Poisson-Konfidenzintervalle für θ, θ_1-θ_2, λ und λ_1-λ_2, Relatives Risiko, Odds-Ratio und zahlreiche andere Verfahren.

Verteilungen
χ^2, t-, F- und Gauß-Verteilung mit Inversen, Nicht-zentrale Verteilungen, Binomial- und Poisson-Verteilung.

Test auf Gauß-Verteilung und Ausreißer
χ^2-Test, Shapiro-Wilk-Test (n beliebig!), Kolmogoroff-Smirnoff-Test, David-Test, Ausreißer-Tests (u.a. Grubbs-Test), Mudholkar-Test auf p-variate Gauß-Verteilung.

Vierfelder- und cxr-Kontingenztafel
χ^2-Test, Fisher-Test (auch zweiseitig-exakt!), McNemar-Test, Median-Test, Cohen's Kappa, Pearson's Kontingenzkoeffizient, Cross-Over-Analyse, Konfidenzintervall für θ_1-θ_2, Mehrdimensionale Tafeln, Vergleich von Tafeln, Bewertung diagnostischer Tests, Äquivalenztest, Mantel-Haenszel-Test, Tests und Konfidenzintervalle für Odds-Ratio und Relatives Risiko, Diagnostische Tests, Terwilliger-Ott- und Bowker's Test und einige andere.

Regressions-und Korrelationsanalyse
Pearson-Regression und Korrelation, Multiple, Polynomiale und Logistische Regression (auch mit Abbau-Modell), Vergleich von Regressionen, Multiple und Partielle Korrelation, Rang-Regression, Spearman's Rang-Korrelation, Residuen-Analyse, Korrelationsmatrix, viele Graphiken.

Parametrische Testverfahren
Ein- und Zweistichproben-t-Test, F-Test, Welch-Test, Ein- und Zweifaktorielle Varianz- und Kovarianzanalyse incl. multipler Vergleiche (Scheffé), Lineare Kontraste, Orthogonale Polynome, Hotelling's Tests, Simultane Konfidenzintervalle, Cross-Over-Analyse (Grizzle), Äquivalenztests.

Nicht-parametrische Testverfahren
Wilcoxon- und Wilcoxon-Mann-Whitney-U-Test, van-Elteren-Test, Tukey- und Moses-Konfidenzintervalle, Hodges-Lehmann- und Wilcoxon-Schätzer, exakte p-Werte, Kolmogoroff-Smirnoff-Test, Kruskal-Wallis-Test incl multipler Vergleiche (Dunn-Holm und Conover), Friedman-Test mit post-hoc-Tests (Willcoxon-Wilcox, Schaich-Hamerle und Conover), Dixon-Mood's Vorzeichen-Test, Test auf Gleichverteilung, Anpassungstest, Cochran's Q-Test, Vergleich von Häufigkeiten bzw Poisson-Verteilungen, Koch's Cross-Over-Analyse, Äquivalenztests, Mantel-Haenszel-Test für k Gruppen und stratifiziert, mit multiplen Vergleichen.

Survival-Analyse
Life-Tables (Kaplan-Meier-Schätzer und Hall-Wellner-Konfidenzbereiche, Graphik), Gehan-Test, Log-Rank-Test von Peto-Pike und Cox-Mantel, Stratifizierter Log-Rank-Test mit Kaplan-Meier-Schätzern, Relativer Hazard, bei k Gruppen mit multiplen Vergleichen, Cox-Regression mit Abbau.

Cluster- und Diskriminanzanalyse
KMeans- und Single-Linkage-Clusteranalyse mit diversen Metriken, Lineare und Quadratische Diskriminanzanalyse, Stoller's nicht-paramametrische Diskriminanzanalyse, Diskriminanzanalyse mit Abbaumodell, Propensity-Score.

Zeitreihen-Analyse
ANOVA, Mann-Test, Iterations- und Phasenhäufigkeitstest, Simultane Tests, Graphiken wie Einzelverläufe und Box- bzw. Bar-Plots der Zeitverläufe.

Faktorenanalyse
Explorative und konfirmatorische Faktorenanalyse mit Zentroid- und Varimax-Methode.

Konfigurationsfrequenzanalyse
Ein- und Mehrstichproben-Konfigurationsfrequenzanalyse mit χ^2-, Binomial- und Lehmacher-Test.

Pharmakologie und Bioäquivalenz
Berechnung u.a. der Größen AUC(t_1,t_n), AUC(t_1,∞), C_{max}, t_{max}, $t_{1/2}$, EC50, Cross-Over-Analyse, Inklusionsregel u.a. mit nicht-parametrischen Tukey- und Moses-Intervallen.

Diagnostische Tests, Methodenvergleich
Falsch-positiv und -negativ, Sensitivität, Spezifität, Effizienz, Prädiktive Werte, Youden-Index (alle mit Konfidenzintervallen), einige Testverfahren, ROC-Kurven mit Test, Bland-Altman- und Passing-Bablok- und Lin-Verfahren mit Tests (alle mit Graphiken).

Toleranzbereiche
Parametrische und nicht-parametrische Toleranzbereiche (uni- und bivariat), Percentilen, Quartilen, Box-Cox-Transformationen, bivariate Bereiche alle mit Graphiken.

Multiple Testverfahren & Meta-Analyse
Multiple Testverfahren nach Bonferroni, Holm, Hommel, Simes und Rüger, Meta-Analyse (Tippett, Fisher, Inverse Normalverteilung), Zwischenauswertungen nach Pocock, O'Brien und Fleming, Bauer-Köhne-Design.

https://www.bias-online.de

Dieses eBook enthält Hyperlinks!

- Ein Klick auf das Penrose-Dreieck der Titelseite führt direkt zum Inhaltsverzeichnis.
- Ein Klick auf einen Abschnitt oder ein Kapitel des Inhaltsverzeichnisses führt zum gewählten Ziel.
- Ein Klick auf die Überschrift eines Abschnitts oder eines Kapitels führt zurück zum Inhaltsverzeichnis.
- „PageDown“ bei Anzeige der Titelseite führt zum Impressum, Vorwort und zum Inhaltsverzeichnis.

Zurück zur Titelseite

Penrose-Dreieck: Lionel S. Penrose, Roger Penrose (1958) Impossible objects: A special type of visual illusion. British Journal of Psychology, Vol. 49 (1), 31-33.